J Biomol NMR (2007) 39:31–52
DOI 10.1007/s10858-007-9175-5

ARTICLE

KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies

Naohiro Kobayashi · Junji Iwahara · Seizo Koshiba · Tadashi Tomizawa · Naoya Tochio · Peter Güntert · Takanori Kigawa · Shigeyuki Yokoyama

Received: 14 March 2007 / Accepted: 15 June 2007 / Published online: 18 July 2007

Abstract The recent expansion of structural genomics has increased the demands for quick and accurate protein structure determination by NMR spectroscopy. The conventional strategy without an automated protocol can no longer satisfy the needs of high-throughput application to a large number of proteins, with each data set including many NMR spectra, chemical shifts, NOE assignments, and calculated structures. We have developed the new software KUJIRA, a package of integrated modules for the systematic and interactive analysis of NMR data, which is designed to reduce the tediousness of organizing and manipulating a large number of NMR data sets. In combination with CYANA, the program for automated NOE assignment and structure determination, we have established a robust and highly optimized strategy for comprehensive protein structure analysis. An application of KUJIRA in accordance with our new strategy was carried out by a non-expert in NMR structure analysis, demonstrating that the accurate assignment of the chemical shifts and a high-quality structure of a small protein can be completed in a few weeks. The high completeness of the chemical shift assignment and the NOE assignment achieved by the systematic analysis using KUJIRA and CYANA led, in practice, to increased reliability of the determined structure.

Keywords NMR · Structure determination · High-throughput · Software · KUJIRA · CYANA · NMRView

Abbreviations

NOESY	Nuclear Overhauser effect spectroscopy
HSQC	Heteronuclear single quantum coherence
SASA	Solvent accessible surface area
TOCSY	Total correlated spectroscopy

Introduction

As demands for protein structure determination by NMR are increasing, especially with the recent expansion in structural genomics, high-throughput techniques are strongly desired. Nowadays, several protocols have been established for the NMR structure determination of moderately sized proteins (smaller than 20 kDa). Considering a structure analysis using modern NMR techniques for a uniformly ^{13}C and ^{15}N labeled protein, the most widely used strategies can be roughly divided into six stages: (i) spectrum data acquisition, (ii) spectrum data processing, (iii) chemical shift assignment of ^{1}H, ^{13}C, and ^{15}N signals, (iv) nuclear Overhauser effect (NOE) assignment and preparation of ^{1}H–^{1}H distance constraints, (v) structure calculation and (vi) validation of the calculated structures.

N. Kobayashi · S. Koshiba · T. Tomizawa ·
N. Tochio · P. Güntert · T. Kigawa · S. Yokoyama (✉)
RIKEN Genomic Sciences Center, 1-7-22, Suehiro-cho,
Tsurumi, Yokohama 230-0045, Japan
e-mail: yokoyama@biochem.s.u-tokyo.ac.jp

J. Iwahara · S. Yokoyama
University of Tokyo, Tokyo 113-0033, Japan

Present Address:
J. Iwahara
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555, USA

T. Kigawa
Tokyo Institute of Technology, Yokohama 226-8502, Japan

In addition to the NOE based distance constraints, dihedral angle constraints and residual dipolar coupling and hydrogen-bonding related constraints are often applied in the structure calculation, if the relevant experimental data sets are available. Although the non-NOE based structural constraints can increase the accuracy of the determined structure, the NOE based distance constraints play a major role in the structure determination process. This is because NOEs are relatively easy to obtain, given their power in structure determination. They thus have a much higher cost performance than other types of NMR data, and are suitable for a high-throughput approach.

The most important data for the structure calculation, ^{1}H–^{1}H distance constraints, are extracted from assignments of NOEs observed in multi-dimensional NOESY spectra. 3D ^{15}N or ^{13}C edited NOESY spectra are preferably used for the extraction of the NOEs, because they presently offer the best compromise between spectral resolution and measurement time needed for their acquisition. The completeness and accuracy of the chemical shift assignments of the ^{1}H, ^{15}N, and ^{13}C signals are crucial for the NOE assignments, and are thus considered first in the NMR analysis. For the NMR chemical shift assignments, a number of 2D and 3D spectrum data sets are normally required, e.g., 4–6 spectra for main-chain signals and 3–4 spectra for side-chain signals. The spectroscopic problems associated with chemical exchange, fast relaxation, severe signal overlap, and/or slight irreproducibility of signal positions among spectra can interfere with achieving complete assignments. Therefore, it is difficult to reach more than 95% accuracy of the chemical shift assignments without any prior knowledge of the correct protein structure. In a structure calculation, distance constraints can be misinterpreted because of missing or erroneous chemical shift data, which lead to distortions in the calculated structure. The next point to consider in an NMR analysis predominantly using NOE derived constraints is the accuracy of the NOE assignments. A total of 4000–6000 NOE peaks and 1500–2000 ^{1}H–^{1}H distance constraints are typically expected from heteronuclear edited NOESY spectra for a ^{15}N- and ^{13}C-labeled protein with a molecular weight around 10 kDa. It is not unusual for three or four ^{1}H, ^{15}N, and ^{13}C signals to be nearly degenerate in a certain place on the 2D HSQC spectrum, and they will be more severely overlapped in the 2D projection of the 3D spectrum data. Due to the degeneracy of the chemical shifts, many NOE peaks will initially be attributed ambiguous assignment possibilities composed of multiple spin pairs. The number of candidates ambiguously assigned is sometimes more than 100, depending on the degeneracy of the signals to be assigned for each spin pair. Including all possible candidates, the number of initial NOE assignments would exceed several tens of thousands. Additionally, small amounts of spectral noise and artifacts unavoidably remain in the large number of NOE peaks in a standard NMR analysis. The interactive effort to iteratively validate, correct, and consolidate such a tremendous number of NOE assignments and to interpret them into distance constraints would occupy a well-experienced NMR scientist for more than a few months.

In order to overcome the obstacles mentioned above, automated or semi-automated methods are the most promising approach to conduct NMR analyses more reliably and with high-throughput. There are many computational programs that attempt automated or semi-automated NOE analysis, as reviewed recently (Altieri and Byrd 2004; Baran et al. 2004; Gronwald and Kalbitzer 2004; Güntert 2003; Nilges and O'Donoghue 1998). The program CYANA (Güntert 2003) includes automated NOE assignment (Herrmann et al. 2002) and structure calculation by torsion angle dynamics (Güntert et al. 1997), and is suitable for high-throughput NMR studies. The new strategies utilized in CYANA, "constraint combination" and "network anchoring" (Herrmann et al. 2002), facilitate finding the correct structure, even if the NOE peak lists contain many artifacts and noise. The reliability of the NOE assignment and the accuracy of the calculated structure have been demonstrated with 90% completeness of the chemical shift assignment (Jee and Güntert 2003). These features of CYANA allow users to concentrate on completing the chemical shift assignment table and the NOE peak table in the early stage of the NMR analysis, rather than on assigning individual NOE cross peaks. In spite of the robustness of the CYANA calculation, especially for the determination of a global chain fold, further completion and refinement of the chemical shift assignments are usually applied to attain higher accuracy in the local regions of the structure. This means that, in the final stage of the NMR analysis, we have to search for the remaining unassigned NOEs and the unidentified NMR signals that may lead CYANA to calculate a structure with a local inaccuracy. Such a structural fault is not critical, but can be enough to confuse a structure-based prediction of the protein function in a future study, if it is located in a putative interaction site of the protein. In the final stage of the analysis, the amount of erroneous data must be extremely small, and a formidable effort is required to eliminate these potential problems. The conventional interactive approaches, such as iteratively scrutinizing the spectrum data sets, the chemical shift table, the NOE peak table and the calculated structure data, can minimize the problem; however, they are time-consuming and greatly reduce the ease of the automated NOE assignment by CYANA, unless there is a smart system to integrate the enormous amount of NMR data. The integration of several analysis tools related to NMR determination into one platform is a promising idea to reduce the tediousness, as they are all based on the concept of standardizing the protocols and the data format. The CCPN data model has been proposed to develop data organization and software pipelines (Fogh et al. 2002, 2005; Vranken et al. 2005). The key feature of the data model is an application programming interface (API) to unify the development of future NMR-related software. The spectral visualization software, ANSIG v3.3, is an example of software development over the CCPN data model (Helgstrand et al. 2000; Kraulis 1989). The SPINS data model has also been proposed, to develop a program suite that can accommodate any type of NMR-related software (Baran et al. 2002). These program suites and systems would work to significantly promote a wide variety of NMR studies if appropriately installed in an NMR laboratory; however, it is still not very easy to introduce them, especially for a non-NMR scientist in a small laboratory who has a good NMR sample and wants a quick analysis.

There are many interactive spectrum analysis programs, e.g., ANSIG (Helgstrand et al. 2000; Kraulis 1989), AURELIA (Neidig et al. 1995), CARA (http://www.nmr.ch), FELIX (FELIX NMR, Inc.), NMRView (Johnson and Blevins 1994), Olivia (http://www.fermi.pharm.hokudai.ac.jp/olivia/), Sparky (Goddard and Kneller 2001), and XEASY (Bartels et al. 1995), which can be used to visualize multidimensional NMR spectra. The distinctive features of NMRView include its capability for directly interpreting a common scripting language, Tcl/Tk, and its high performance in quickly drawing contour plots of 2D planes from multi-dimensional spectra. Relying on these features, several software add-ons, such as Smartnotebook (Slupsky et al. 2003) and NvAssign (Kirby et al. 2004), have been released.

Here, we have developed a new software package, called KUJIRA, consisting of integrated modules for the systematic and interactive analysis of NMR data. NMRView is used together with KUJIRA for controlling spectrum windows and for executing external Tcl/Tk source code and C programs. The modules of KUJIRA are designed to simultaneously accommodate a large number of spectrum data sets, chemical shift data, NOE peak data, NOE assignment data, and structure data, which are systematically organized (Fig. 1). One module in KUJIRA can import output files from an automated NOE assignment performed by CYANA and control the spectrum windows based on the NMR parameters related to the NOE peaks. Another module of KUJIRA is used to assess the structural quality of the calculated structures. It facilitates the analysis of the geometrical features of the calculated structure, i.e., the solvent accessible surface area, the distributions of the dihedral angles (ϕ, ψ), and (χ^1, χ^2), and the secondary structure elements. The modules of KUJIRA seamlessly control the chemical shift tables, the spectrum windows and the NOE assignment tables, so that users can intuitively and expeditiously address artifact or noise peaks in the NOE peak table as well as misassignments in the chemical shift table. These functions in KUJIRA allow users not only to obtain highly accurate chemical shifts but also to save a considerable amount of time during the structure refinement. In this article, we will describe the construction of the NMR program suite KUJIRA, and the functional modules in KUJIRA to apply the

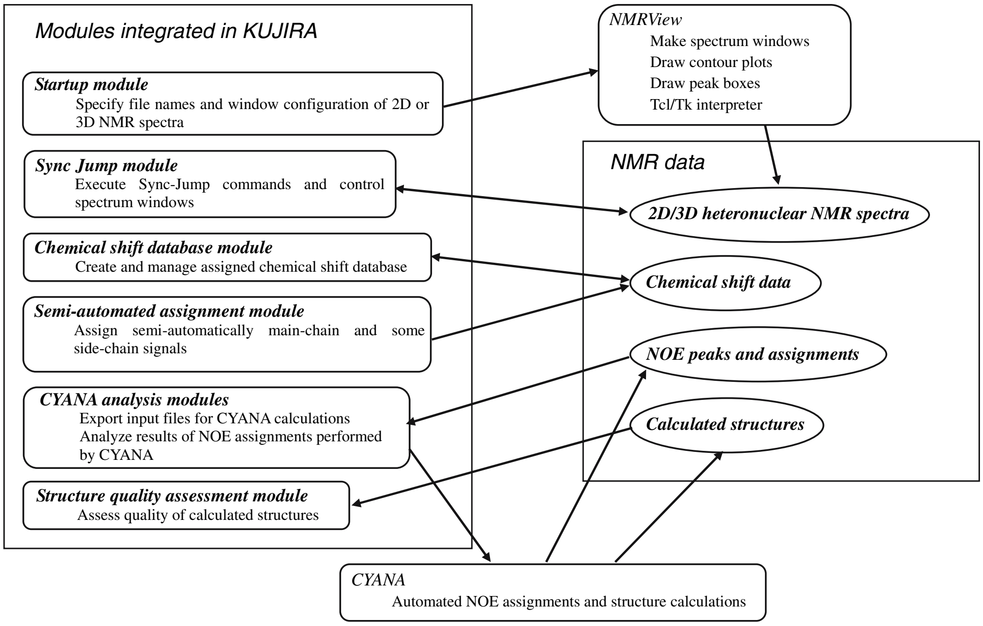

Modules integrated in KUJIRA
Startup module
Specify file names and window configuration of 2D or 3D NMR spectra
Sync Jump module
Execute Sync-Jump commands and control spectrum windows
Chemical shift database module
Create and manage assigned chemical shift database
Semi-automated assignment module
Assign semi-automatically main-chain and some side-chain signals
CYANA analysis modules
Export input files for CYANA calculations
Analyze results of NOE assignments performed by CYANA
Structure quality assessment module
Assess quality of calculated structures
NMRView
Make spectrum windows
Draw contour plots
Draw peak boxes
Tcl/Tk interpreter
NMR data
2D/3D heteronuclear NMR spectra
Chemical shift data
NOE peaks and assignments
Calculated structures
CYANA
Automated NOE assignments and structure calculations

Fig. 1 KUJIRA modules and other software used for the NMR structure studies. The scheme illustrates how the modules in KUJIRA interact with other software and NMR data used for NMR analyses (see details in the text). All of the modules in KUJIRA can share the loaded NMR data, such that the user can interactively control the spectrum windows, based on the chemical shift and NOE assignments

semi-automated NMR signal assignments required for the fully automated NOE assignment and structure calculation program CYANA, as well as a strategy specifically directed to high-throughput NMR structure studies, using the functions of KUJIRA in conjunction with CYANA.

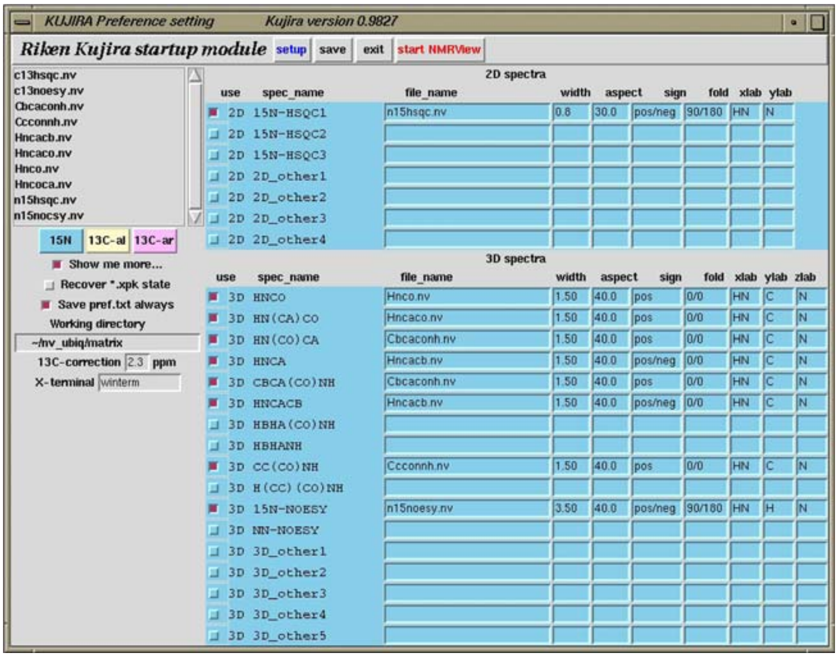
Materials and methods

Construction of the software, KUJIRA

The scriptable and graphics-based program NMRView is used to display and control the 2D and 3D spectrum contour plots. The C programming language version of NMRView, available from One Moon Scientific Inc. (http://www.onemoonscientific.com), works on a variety of computer platforms. The intrinsic capability of NMRView to load and interpret external Tcl/Tk sources is utilized in the KUJIRA package to build the graphical user interface (GUI) and for the numerous subroutines that implement the functions of KUJIRA. In addition, the external C programs included in the KUJIRA package are accessed from the subroutines to achieve a high performance for computation-intensive numerical calculation jobs. The KUJIRA modules described in this paper are shown in the scheme of Fig. 1, which illustrates how the modules in KUJIRA interact with other software and with the NMR data used for the structure analysis.

Startup module

The graphical interface of the Startup module, shown in Fig. 2, appears first when the user starts KUJIRA. Using the interface, the user can readily specify the file names and the directory path names of the spectrum data sets, as well as the parameters for drawing their contour plots. Once the user launches NMRView from the startup module, the graphical interfaces and subroutines inherently implemented in NMRView are transiently created. Some of the menus of graphical interfaces and subroutines are suppressed by the Startup module of KUJIRA, and then this is followed by the construction of the other KUJIRA graphical interfaces and subroutines for making networks linked to the spectrum windows, chemical shift data, NOE assignment data and structure data. After these preliminary jobs, the NMRView startup is completed by restructuring the spectrum windows and the graphical interfaces of the KUJIRA modules, according to the preferences specified in the Startup module.



c13hsqc.nv
c13noesy.nv
Cbcaconh.nv
Ccconnh.nv
Hncacb.nv
Hncaco.nv
Hnco.nv
Hncoca.nv
n15hsqc.nv
n15nocsy.nv

☒ Show me more...
□ Recover *.xpk state
☒ Save pref.txt always
Working directory
~/nv_ubiq/matrix
13C-correction 2.3 ppm
X-terminal winterm

2D spectra

use	spec_name	file_name	width	aspect	sign	fold	xlab	ylab
☒	2D 15N-HSQC1	n15hsqc.nv	0.8	30.0	pos/neg	90/180	HN	N
□	2D 15N-HSQC2							
□	2D 15N-HSQC3							
□	2D 2D_other1							
□	2D 2D_other2							
□	2D 2D_other3							
□	2D 2D_other4							

3D spectra

use	spec_name	file_name	width	aspect	sign	fold	xlab	ylab	zlab
☒	3D HNCO	Hnco.nv	1.50	40.0	pos	0/0	HN	C	N
☒	3D HN(CA)CO	Hncaco.nv	1.50	40.0	pos	0/0	HN	C	N
☒	3D HN(CO)CA	Cbcaconh.nv	1.50	40.0	pos	0/0	HN	C	N
☒	3D HNCA	Hncacb.nv	1.50	40.0	pos/neg	0/0	HN	C	N
☒	3D CBCA(CO)NH	Cbcaconh.nv	1.50	40.0	pos	0/0	HN	C	N
☒	3D HNCACB	Hncacb.nv	1.50	40.0	pos/neg	0/0	HN	C	N
□	3D HBHA(CO)NH								
□	3D HBHANH								
☒	3D CC(CO)NH	Ccconnh.nv	1.50	40.0	pos	0/0	HN	C	N
□	3D H(CC)(CO)NH								
☒	3D 15N-NOESY	n15noesy.nv	3.50	40.0	pos/neg	90/180	HN	H	N
□	3D NN-NOESY								
□	3D 3D_other1								
□	3D 3D_other2								
□	3D 3D_other3								
□	3D 3D_other4								
□	3D 3D_other5								

Fig. 2 The graphical interface of the Startup module in KUJIRA. Prior to the analysis, the user must specify the spectrum data sets and several parameters in this module. The small list-box on the upper left of the interface displays the currently available spectrum data sets, from which the user may select one to transfer into the entries of the corresponding spectrum type on the right side of the interface. The parameters to configure a narrow region of 2D spectra can also be specified on the entries arrayed on the same spectrum entry row. The check-buttons, on the left of the spectrum name labels, are used to select the spectra that are being loaded for the current analysis. The buttons just below the list-box convert the interface mode between the different "Sync-Jump" types: ^{15}N-, ^{13}C-aliphatic, and ^{13}C-aromatic

Sync-Jump module

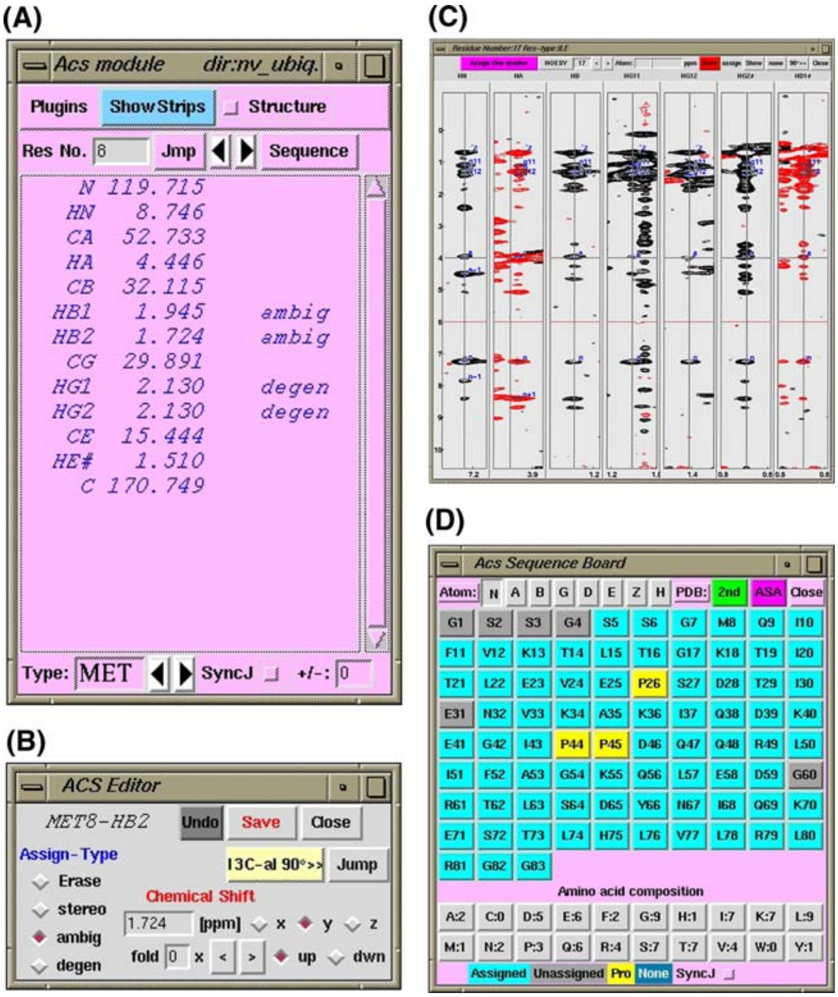
The synchronized spectrum jump system, referred to as "Sync-Jump", is the most important function of KUJIRA. Each of the loaded spectrum data sets is attributed to one of the Sync-Jump classes, as specified in the startup module. The Sync-Jump commands synchronize and simultaneously display groups of 2D spectrum strips extracted from 3D spectra attributed to the same Sync-Jump class, while concurrently changing their drawing position in the 2D HSQC projection. The 2D HSQC spectra are similarly controlled by the Sync-Jump command to display the narrow region centering on the currently specified position. As the default setting, three Sync-Jump classes are available, namely ^{15}N- and ^{13}C-aliphatic and ^{13}C-aromatic, respectively corresponding to the Sync-Jump action on the 2D ^{1}H–^{15}N, ^{1}H–^{13}C for the aliphatic and aromatic signals-correlated spectrum projection of the 3D spectrum. All modules in KUJIRA provide Sync-Jump commands, so the user can expeditiously manipulate the number and contents of the spectrum windows shown. There are two ways to provide chemical shift information to the Sync-Jump command. In the first way, the command is simply executed with the given chemical shifts of a ^{1}H and a heavy atom dimension. For instance, this function works by clicking on the 2D-HSQC type spectrum window in the "Click-and-Jump mode", which gives the chemical shifts at the clicked position, and then triggers the simultaneous display of the 2D spectrum strips extracted from the 3D spectra attributed to the same Sync-Jump class. During NOE peak analysis, this protocol is also helpful to manipulate the 3D NOESY spectra as well as the other 3D spectra used for fine adjustments of the assigned chemical shifts. The second Sync-Jump protocol is based on groups of atoms with assigned chemical shifts, and requires the chemical shift assignments of both the protons and heavy atoms of interest. Using this protocol, Sync-Jump commands are provided to the Chemical shift database module, the module for the analysis of NOE assignments, and the module for the structure quality assessment, as described below. The subroutines responsible for the "Sync-Jump" functionality are integrated in the Sync-Jump module of KUJIRA.

Chemical shift database module

The Chemical shift database module of KUJIRA is used to manage the database of experimentally determined chemical shifts for the signals of proton, nitrogen, and carbon atoms. The module provides several graphical interfaces (Fig. 3). One of them, the main graphical interface shown in Fig. 3A, can display the assigned chemical shift data for a desired amino acid residue. The pop-up window from the main graphical interface shown in Fig. 3B allows the user to interactively input and correct the chemical shifts directly from the spectrum at any axis, by clicking on the spectrum window. The Chemical shift database module provides a function to automatically examine the validity of the defined chemical shifts, by comparing them with the chemical shift statistics for proteins derived from the BMRB database (http://www.bmrb.wisc.edu/). Any anomalous chemical shift can immediately be highlighted with an alert message in the main graphical interface. Since, this module provides the Sync-Jump command, to control the strips from 3D spectra based on the values in the chemical shift table, the user is able to confirm and correct the signal assignments for each residue by directly observing the related 2D strips. The window shown in Fig. 3C is used for simultaneously confirming all of the chemical shifts of the main-chain and side-chain signals of the currently analyzed residue. In the graphical interface, 2D spectrum strips extracted from 3D NOESY or TOCSY spectra are arrayed to display narrow regions corresponding to the spectrum positions of the assigned main-chain and side-chain signals. The cross-peaks based on the assignment of intra-residual signals (and sequential signals for NOESY) are indicated with the blue boxes in each 2D spectrum strip. The most powerful function of this interface is that the user can intuitively notice an inaccuracy in the defined chemical shifts, by monitoring the expected intra-residual and sequential NOEs in the arrayed spectrum strips, thus helping to interactively update the chemical shifts by clicking on the appropriate spectrum position. Another pop-up graphical interface (Fig. 3D) displays the amino acid sequence of the analyzed protein and informs the user about the completeness of the assignments by color-coded buttons.

Semi-automatic assignment module

KUJIRA has a module for the semi-automated sequence-specific assignment of the main-chain signals, based on the ^{13}C$_{\alpha}$, ^{13}C$_{\beta}$, and ^{13}C′ signals derived from the triple resonance spectra, including several graphical interfaces (see Figs. 4–6) and a key external C program, "QuickAssign." The QuickAssign program performs the assignment jobs that are described in Fig. 5, based on the basic main-chain assignment information, such as the peak IDs and chemical shifts of ^{1}H$_{N}$, ^{15}N$_{\alpha}$ and the sequential and intra-residual ^{13}C$_{\alpha}$, ^{13}C$_{\beta}$, and ^{13}C′ signals. Prior to running the program, the user has to prepare the information manually or automatically. There is a tool for creating a text file listing the identified peak IDs, corresponding to the ^{1}H$_{N}$–^{15}N$_{\alpha}$ signal correlations from the peak table file of 2D or 3D heteronuclear correlation spectra, such as ^{1}H–^{15}N HSQC, 3D HNCO, and 3D-HN(CO)CA. The user is allowed to



```
 N 119.715
HN   8.746
CA  52.733
HA   4.446
CB  32.115
HB1  1.945  ambig
HB2  1.724  ambig
CG  29.891
HG1  2.130  degen
HG2  2.130  degen
CE  15.444
HE#  1.510
 C 170.749
```

G1	S2	S3	G4	S5	S6	G7	M8	Q9	I10
F11	V12	K13	T14	L15	T16	G17	K18	T19	I20
T21	L22	E23	V24	E25	P26	S27	D28	T29	I30
E31	N32	V33	K34	A35	K36	I37	Q38	D39	K40
E41	G42	I43	P44	P45	D46	Q47	Q48	R49	L50
I51	F52	A53	G54	K55	Q56	L57	E58	D59	G60
R61	T62	L63	S64	D65	Y66	N67	I68	Q69	K70
E71	S72	T73	L74	H75	L76	V77	L78	R79	L80
R81	G82	G83							

A:2	C:0	D:5	E:6	F:2	G:9	H:1	I:7	K:7	L:9
M:1	N:2	P:3	Q:6	R:4	S:7	T:7	V:4	W:0	Y:1

Fig. 3 (**A**) The main graphical interface of the Chemical shift database module. This interface displays the assigned chemical shift data for a desired amino acid residue. Each row in the list-box indicates the atom name, the assigned chemical shift value, and the ambiguity status of the assigned signal; "ambig" means ambiguously assigned, "degen" means the signals are degenerate, and "stereo" means stereo-specifically assigned. (**B**) Pop-up graphical interface from the main graphical interface for the proton signal of Met8-Hβ2. The assigned chemical shift is displayed in the center of the interface. The user may edit the assigned chemical shift or the ambiguity status of the signal. The spectrum strips associated with this atom can readily be accessed from this interface by Sync-Jump commands. (**C**) Pop-up graphical interface from the main graphical interface, for the confirmation and correction of the main-chain and side-chain signal assignments. It displays the relevant 2D spectrum strips extracted from the 3D-NOESY spectra. The arrayed 2D spectrum strips show narrow regions corresponding to the assigned ^{1}H–^{15}N or ^{1}H–^{13}C signal positions. The expected intra-residue and sequential-cross peaks are indicated by blue boxes, labeled with the assignment of the signals in the indirect dimension using simplified nomenclature, such as *n*: HN(*i*), *n* – 1: HN(*i* – 1), *a*: Hα(*i*) and so on. The chemical shifts can be interactively adjusted by mouse operations. (**D**) Pop-up graphical interface from the main graphical interface that represents the amino acid sequence of the currently analyzed protein. The labels on the buttons indicate the one-letter amino acid code and the residue number. The buttons are used to switch the information displayed in the main graphical interface of the Chemical shift database module

identify the sequential and intra-residual $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and $^{13}C'$ signals, which can be observed on the indirect dimension of the ^{1}H-^{15}N position specified by the peak ID from the 3D triple resonance spectra and a graphical interface (Figs. 4A, B), either manually or automatically. After the preparation of the text file describing the basic main-chain information, the user can run the automated sequence-specific assignment C program, "QuickAssign", to search for the segments of peak IDs. QuickAssign quickly performs automated sequential assignments based on the semi-exhaustive search algorithm, as shown in the program flow chart (see Fig. 5). The program makes a non-redundant list of all possible segments, by finding and linking the peak IDs based on the sequential connectivity of the identified $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and $^{13}C'$ signals, using the specified error tolerance (default values are set at 0.3 ppm for all atom types). QuickAssign maps the segment *x* at the residue number *k* on the amino acid sequence of the sample protein, and evaluates the penalty value *P*(*x*,*k*) based on the following Eqs. 1 and 2.

$$P(x,k) = \frac{\sum_{i=1}^{N(x)} X^{\alpha\beta}(x,k,i)}{N(x)} \tag{1}$$

where $X^{\alpha\beta}(x,k,i)$ is the probability value for the segment x at the i-th residue for $^{13}C_{\alpha}/^{13}C_{\beta}$, which is given by:

$$t_i = \left(\frac{A_i^{\alpha} - \bar{A}^{\alpha}_{R(k+i-1)}}{\delta^{\alpha}_{R(k+i-1)}}\right)^2 + \left(\frac{A_i^{\beta} - \bar{A}^{\beta}_{R(k+i-1)}}{\delta^{\beta}_{R(k+i-1)}}\right)^2 \tag{2}$$

$$\begin{cases} X^{\alpha\beta}(x,k,i) = t_i & \text{if } i = 1 \text{ or } (i > 1,\ t_i \le t_{i-1}) \\ X^{\alpha\beta}(x,k,i) = t_{i-1} & \text{if } (i > 1,\ t_i > t_{i-1}) \end{cases}$$

where A_i^{α} and A_i^{β} are the chemical shifts for the identified $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ signals, respectively, and $\bar{A}^{\alpha/\beta}_{R(k+i-1)}$ and $\delta^{\alpha/\beta}_{R(k+i-1)}$ are the average and standard deviation, respectively of the $^{13}C_{\alpha}/^{13}C_{\beta}$ signal corresponding to amino acid type R at residue $k + i - 1$. The average and standard deviation are derived from the restricted statistics calculated for the amino acid specific $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts in the BMRB database.

The mapped segment will be judged by the following criteria: (1) the mapped segment on the residue number k should have the best penalty value *P(x,k)* three times as low as the secondary one, (2) the penalty value must be below the specified maximal value, *Maxp*, (3) the segment is longer than the specified minimal length, *Minl*, (4) it does not include any sequential connectivity with the amount of used chemical shift information below the minimal value, *Minc* and (5) it does not include ambiguous connectivities greater than the specified value, *Amb*. If the mapped segment satisfies all of these criteria, then the N- and C-terminal residues of the segment are truncated, and the peak IDs included in the segment will be suppressed in the subsequent assignment step. If the longer segment contains the identical peak ID sequence of the shorter one and they satisfy the criteria, then the longer one is chosen. After four rounds of iterative assignment steps, the best-mapped segments are displayed on the graphical interface along with the calculated penalty values, as shown in Fig. 4E and F. If the user finds that the segments are correctly mapped, then the user can assign the segment permanently on the graphical module. The performance of the program, from loading the assignment data to displaying the best results of mapping the segments, depends on the number of identified peak IDs and the length of the sample protein. For a 100–150 residue protein, the calculation typically takes 0.5–1.0 s on a standard PC or workstation. The completeness of the assignments is around 70–80% using the default setting parameters, as in the example shown in Table 1, and it also depends on the completeness of the user-prepared information for the main-chain assignment. If there are many missing signals, minor peaks, artifact peaks, some repeated sequence in the protein sequence, it will be very difficult to achieve the complete assignment of the main-chain signals by only using QuickAssign.

Since the module achieves the sequence-specific assignments of the main-chain signals using the above-mentioned method based on well-separated peak IDs, the assignments may fail if many peak IDs of the $^{1}H_{N}$–$^{15}N_{\alpha}$ signals are missing, because of severe overlapping. In this case, the search function on the main graphical module is helpful to discover the potential peaks. The module can search for the sequential residue at the (i – 1) or (i + 1) position by calculating the values from the spectrum intensity of 3D triple resonance spectra, using the identified chemical shifts of $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and $^{13}C'$ signals, as described in Fig. 4B.

The sequential connectivities of the assigned peak IDs can be confirmed on a graphical interface (Figs. 6A, B), where the 2D spectrum strips extracted from the 3D spectra corresponding to the peak ID position of the assigned segments are arrayed, and the identified ^{13}C signal positions are indicated with the blue boxes. There is an assignment assessment tool, "CheckAssign", which can be used to find problems in the sequence-specific assignments. It allows the user to correct them interactively on a graphical interface, as shown in Figs. 6C and D.

After the semi-automated main-chain signal assignments, the user can export the assigned chemical shifts to the chemical shift database. In addition to the data exportation, the user is allowed to assign the $^{1}H_{\alpha}$ and $^{1}H_{\beta}$ signals, if 2D ^{1}H–^{13}C HSQC and 3D HBHA(CBCACO)NH spectra are available. The external C program automatically picks the peaks on 2D ^{1}H–^{13}C HSQC and 3D HBHA(CBCACO)NH, then tries to find peaks in the spectrum region corresponding to the assigned $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ signals. Only the signals with corresponding $^{1}H_{\alpha}$–$^{13}C_{\alpha}$, and $^{1}H_{\beta}$–$^{13}C_{\beta}$ peaks found on both 2D ^{1}H–^{13}C HSQC and 3D HBHA(CBCACO)NH are assigned.

CYANA analysis modules

KUJIRA provides an interactive module to analyze the results of the automated NOE assignment by CYANA. The user can import a log file of the NOE assignments and inspect the results for each peak, including all of the assignment candidates, in the graphical interface shown in Fig. 7A. A variety of GUI features are implemented in this module: the increment and decrement buttons used for stepping through the assignment results with respect to the peak ID number, the menu buttons to switch between NOE peak lists, and between assigned and unassigned NOEs, and the checkboxes to skip NOE peaks that may not need to be inspected, such as those without or with small

Fig. 4 (**A**) 2D spectrum strips extracted from the 3D triple resonance spectra, which can be used for the identification of ^{13}C signals for sequence-specific main-chain signal assignments. By the "Sync-Jump" function of KUJIRA, the spectrum strips synchronously jump to the ^{1}H$_{N}$–^{15}N$_{\alpha}$ position as their class is specified as "^{15}N". (**B**) The function to predict the amino acid type for the sequential (i – 1) and current peak ID (i) can be carried out by pressing the button "TellMeResTP", using the chemical shifts, $\mathrm{A}_{\mathrm{i}}^{\alpha}$ and $\mathrm{A}_{\mathrm{i}}^{\beta}$, for the identified sequential or intra-residual ^{13}C′$_{\alpha}$ and ^{13}C$_{\beta}$ signals. The probability value $P(i,R)$ for a amino acid type R is calculated by the following equation:

$$p = \left(\frac{A_i^{\alpha} - \bar{A}_R^{\alpha}}{\delta_R^{\alpha}}\right)^2 + \left(\frac{A_i^{\beta} - \bar{A}_R^{\beta}}{\delta_R^{\beta}}\right)^2 \qquad (3)$$

$$P(i,R) = \frac{1}{2\pi\delta_R^{\alpha}\delta_R^{\beta}}\exp\left(-0.5p\right)$$

where $\bar{A}_R^{\alpha/\beta}$ and $\delta_R^{\alpha/\beta}$ are the average and standard deviation, respectively, of the ^{13}C$_{\alpha}$/^{13}C$_{\beta}$ signal corresponding to amino acid type R. These values are derived from the restricted statistics calculated for the amino acid specific ^{13}C$_{\alpha}$ and ^{13}C$_{\beta}$ chemical shifts in the BMRB database. The prediction function calculates the value, $P(i,R)$, for all amino acid types, sorts them by the values in increasing order, and then displays the amino acid types with the value greater than 10^{-4} in the one-letter code in the entry widgets, as indicated with the red boxes in (**B**). The function to search for the sequential peak IDs on the positions (i – 1) and (i + 1) can be carried out by pressing the "Search" button in the middle of the interface. The function calculates the sum of the spectrum intensity $I(i,j)$ on the xz-positions specified with the chemical shifts of the ^{1}H$_{N}$ and ^{15}N$_{\alpha}$ signals for the target peak ID j and on the y-position specified with the chemical shifts of the sequential or intra-residual ^{13}C$_{\alpha}$, ^{13}C$_{\beta}$ and ^{13}C′ signals of the current peak ID i.

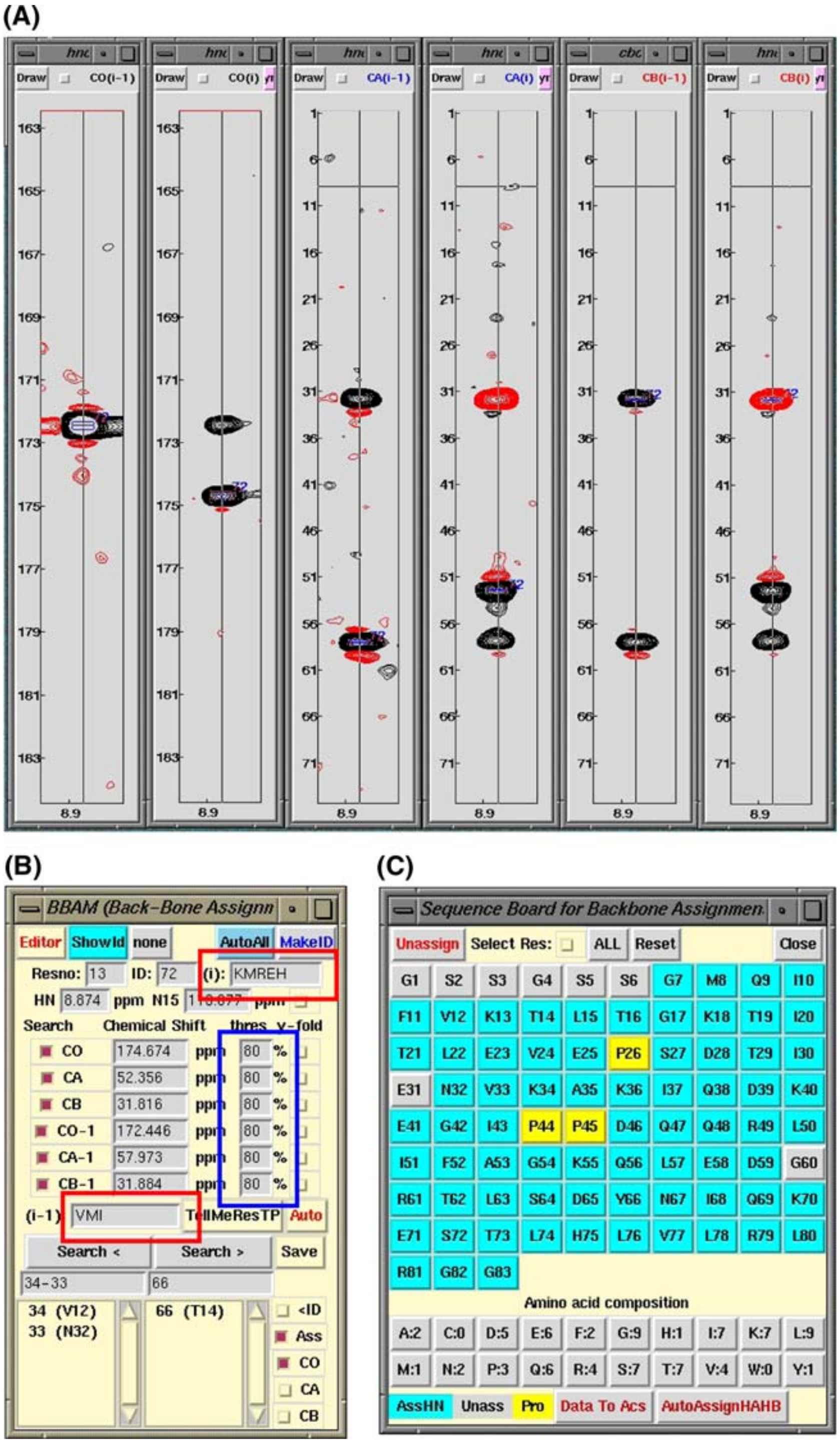
$$I(i,j) = k_{\mathrm{CO}}\left|I^{\mathrm{CO}}\right| + k_{\alpha}\left|I^{\alpha}\right| + k_{\beta}\left|I^{\beta}\right| \qquad (4)$$

where the values k_{CO}, k_{α}, and k_{β} are normalization factors (default values are 1.0) for the detected intensity, I^{CO}, I^{α}, and I^{β} of the spectra. If the calculated value $I(i,j)$, multiplied by the factor specified on the entrance widgets of the graphical interface (indicated by the blue box), is below the specified threshold for each 3D spectrum; then the value will be zero. The scanned peak IDs are sorted by the calculated values and displayed on the list-box at the bottom of the graphical interface, if the value is greater than zero. By double-clicking on one of the items in the list-box, a spectrum strip window corresponding to the scanned peak ID will appear. (**C**) Pop-up graphical interface, displaying the amino acid sequence of the sample protein. The assigned residues are colored by cyan. By pressing the "Data to Acs" button on the bottom of the interface, the user can export the assigned chemical shifts to the assignment database module. The "Auto-AssignHAHB" button can be used to run the data exportation and the fully-automated assignment of the ^{1}H$_{\alpha}$ and ^{1}H$_{\beta}$ signals using 2D ^{1}H–^{13}C HSQC and HBHA(CBCACO)NH spectra. (**D**) Pop-up graphical interface used for managing the main-chain assignment information, such as the peak ID, the chemical shifts of ^{1}H$_{N}$, ^{15}N$_{\alpha}$, and the sequential and intra-residual ^{13}C$_{\alpha}$, ^{13}C$_{\beta}$, and ^{13}C′ signals. The buttons labeled with the name of the item can be used to sort the peak ID information by the corresponding item type. (**E**) Pop-up graphical interface used to run the external C program "QuickAssign" (see text). The user can temporarily adjust all of the ^{13}C chemical shifts of the main-chain data, only for the functions of the modules, "QuickAssign" and "CheckAssign", by specifying the correction value of the entry in the interface. The tolerance values for making peak ID segments can also be adjusted in the entries on the interface (default values are 0.3 ppm for all atom types). (**F**) By double-clicking one of the mapped segments in the interface of (**E**), a graphical interface will pop-up to display the detail of the mapped segment. The user can decide to permanently assign the mapped segment by pressing the button "Assign" on the top of the interface ▶

associated violations of distance constraints. Together with the Sync-Jump command, the user can easily and quickly access the desired NOE peaks for inspection. The module has a function to display the 2D spectrum strip that is orthogonally transposed to the indirect dimension of the NOE assignment side-by-side with the spectrum strip of the current NOE peak, as shown in Fig. 7B. For a NOE with multiple ambiguous assignments, the user can switch to the symmetrical spectrum strip corresponding to another possible assignment candidate, by double-clicking it in the interface shown in Fig. 7A. In another mode of the interface (Fig. 7C), the NOE peaks can be filtered and sorted with respect to various parameters, such as the prime assignment candidate, the peak number, the peak intensity, the ^{1}H–^{1}H upper distance bound, the observed violation, or the peak position on the x-, y-, or z-axis or in the xz-plane. The module has a function to evaluate the symmetry of the NOE peak assignments for which the intensity at the symmetrical position defined by the most probable NOE assignment is measured in the spectrum. If this intensity is greater than a user-specified threshold, then the peak entry is colored blue; otherwise, it is red. Another graphical interface is available for setting up a CYANA calculation (data not shown), in which the user can specify the version of CYANA and the path names of the NOE peak tables, and then save these settings for later use by the CYANA analysis module to properly interpret the CYANA output files. The interface can be used to prepare a directory for a CYANA calculation, and to create a sequence file, a chemical shift table, NOE peak tables and an initial settings file (init.cya) in the format required by CYANA.

Structure quality assessment module

The KUJIRA module for the quality assessment of calculated protein structures is implemented as an external C-program. The module can load a set of ensembled structures that are formatted in a variety of coordinate types, including DYANA and PDB file formats. The secondary structure elements are detected by the Kabsch and Sander method (Kabsch and Sander 1983), and the percentage of structures is evaluated for each residue in which the secondary structure is detected. The types of secondary structures with a percentage greater than 30% are classified



Resno: 13 ID: 72 (i): KMREH

HN 8.874 ppm N15 1[illegible] ppm □

Search	Chemical Shift		thres		y-fold
■ CO	174.674	ppm	80	%	□
■ CA	52.356	ppm	80	%	□
■ CB	31.816	ppm	80	%	□
■ CO-1	172.446	ppm	80	%	□
■ CA-1	57.973	ppm	80	%	□
■ CB-1	31.884	ppm	80	%	□

(i-1) VMI

34-33 66

34 (V12)
33 (N32)

66 (T14)

□ <ID
■ Ass
■ CO
□ CA
□ CB

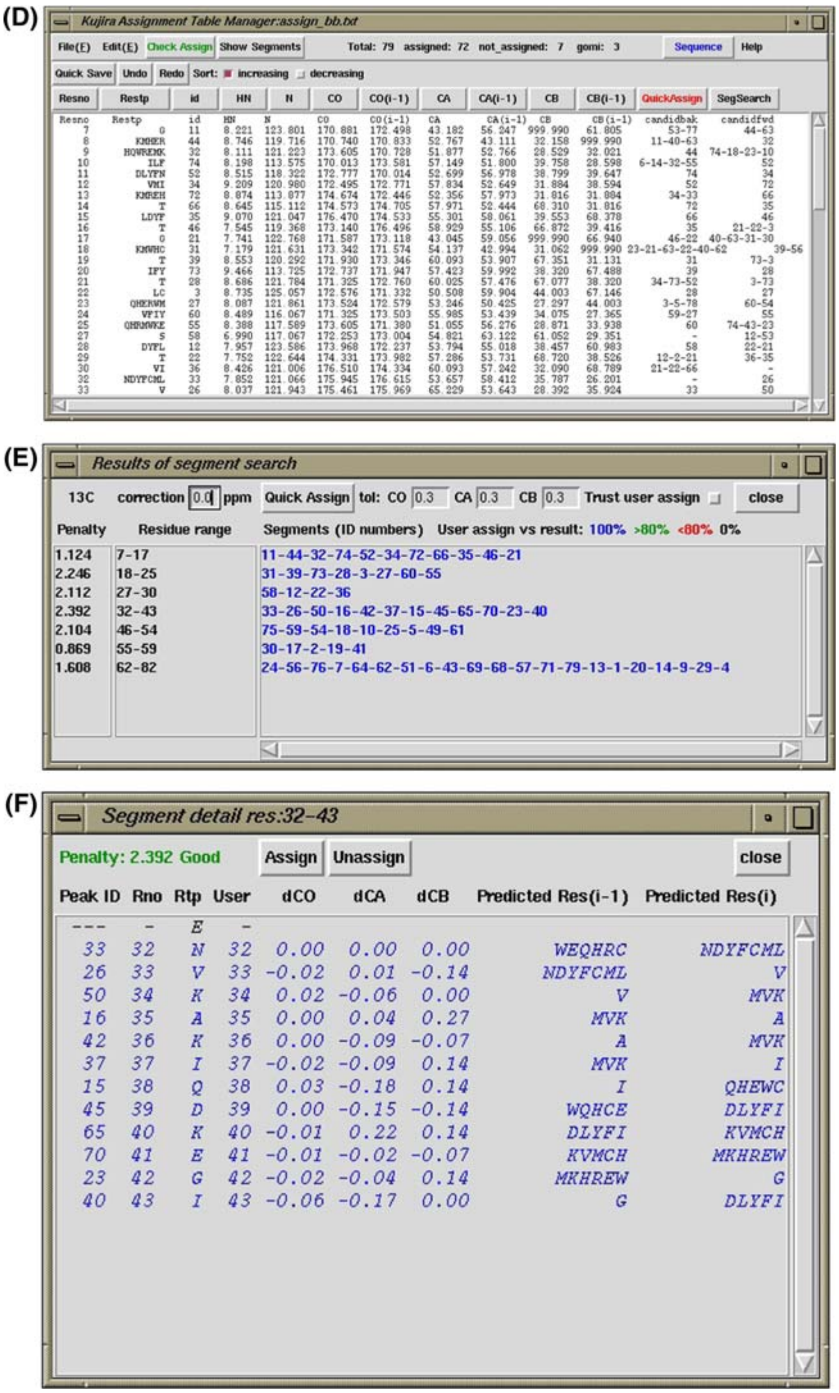
Select Res: □

G1	S2	S3	G4	S5	S6	G7	M8	Q9	I10
F11	V12	K13	T14	L15	T16	G17	K18	T19	I20
T21	L22	E23	V24	E25	P26	S27	D28	T29	I30
E31	N32	V33	K34	A35	K36	I37	Q38	D39	K40
E41	G42	I43	P44	P45	D46	Q47	Q48	R49	L50
I51	F52	A53	G54	K55	Q56	L57	E58	D59	G60
R61	T62	L63	S64	D65	Y66	N67	I68	Q69	K70
E71	S72	T73	L74	H75	L76	V77	L78	R79	L80
R81	G82	G83							

Amino acid composition

A:2	C:0	D:5	E:6	F:2	G:9	H:1	I:7	K:7	L:9
M:1	N:2	P:3	Q:6	R:4	S:7	T:7	V:4	W:0	Y:1

and displayed in the main graphical interface of the module (see Fig. 8A). The solvent accessible surface area for each amino acid residue is calculated, using the method established by Lee and Richards (1971). If the average solvent accessible surface area of the side-chain atoms is smaller than 30%, then the residue is judged to be buried inside the protein, as shown in Fig. 8B. The values of the dihedral angles ϕ, ψ, χ^1, and χ^2 are also calculated by the module. If the values of a dihedral angle are clustered in two or more distinct conformations, then the module displays a caution message and reports details about the split dihedral angle conformation. The module can further perform a Ramachandran plot analysis for the main-chain angles (ϕ,ψ), according to the methods established and applied in the program PROCHECK (Laskowski et al. 1996) (Fig. 8A), and a rotamer analysis for the side-chain dihedral angles


(D)
Kujira Assignment Table Manager:assign_bb.txt
File(F) Edit(E) Check Assign Show Segments
Sequence Help
Quick Save Undo Redo
Resno Restp id HN N CO CO(i-1) CA CA(i-1) CB CB(i-1) QuickAssign SegSearch

Total: 79 assigned: 72 not_assigned: 7 gomi: 3

Sort: ■ increasing □ decreasing

```
Resno  Restp     id  HN     N        CO       CO(i-1)  CA      CA(i-1)  CB       CB(i-1)  candidbak          candidfwd
   7         G   11  8.221  123.801  170.881  172.498  43.182  56.247   999.990   61.805             53-77        44-63
   8     KMHER   44  8.746  119.716  170.740  170.833  52.767  43.111    32.158  999.990          11-40-63           32
   9   HQWREMK   32  8.111  121.223  173.605  170.728  51.877  52.766    28.529   32.021                44  74-18-23-10
  10       ILF   74  8.198  113.575  170.013  173.581  57.149  51.800    39.758   28.598        6-14-32-55           52
  11     DLYFN   52  8.515  118.322  172.777  170.014  52.699  56.978    38.799   39.647                74           34
  12       VMI   34  9.209  120.980  172.495  172.771  57.834  52.649    31.884   38.594                52           72
  13     KMREH   72  8.874  113.877  174.674  172.446  52.356  57.973    31.816   31.884             34-33           66
  14         T   66  8.645  115.112  174.573  174.705  57.971  52.444    68.310   31.816                72           35
  15      LDYF   35  9.070  121.047  176.470  174.533  55.301  58.061    39.553   68.378                66           46
  16         T   46  7.545  119.368  173.140  176.496  58.929  55.106    66.872   39.416                35      21-22-3
  17         G   21  7.741  122.768  171.587  173.118  43.045  59.056   999.990   66.940             46-22  40-63-31-30
  18     KMWHC   31  7.179  121.631  173.342  171.574  54.137  42.994    31.062  999.990 23-21-63-22-40-62            39-56
  19         T   39  8.553  120.292  171.930  173.346  60.093  53.907    67.351   31.131                31         73-3
  20       IFY   73  9.466  113.725  172.737  171.947  57.423  59.992    38.320   67.488                39           28
  21         T   28  8.686  121.784  171.325  172.760  60.025  57.476    67.077   38.320          34-73-52         3-73
  22        LC    3  8.735  125.057  172.576  171.332  50.508  59.904    44.003   67.146                28           27
  23    QHERWM   27  8.087  121.861  173.524  172.579  53.246  50.425    27.297   44.003            3-5-78        60-54
  24      VFIY   60  8.489  116.067  171.325  173.503  55.985  53.439    34.075   27.365             59-27           55
  25   QHRMWKE   55  8.388  117.589  173.605  171.380  51.055  56.276    28.871   33.938                60     74-43-23
  27         S   58  6.990  117.067  172.253  173.004  54.821  63.122    61.052   29.351                 -        12-53
  28      DYFL   12  7.957  123.586  173.968  172.237  53.794  55.018    38.457   60.983                58        22-21
  29         T   22  7.752  122.644  174.331  173.982  57.286  53.731    68.720   38.526           12-2-21        36-35
  30        VI   36  8.426  121.006  176.510  174.334  60.093  57.242    32.090   68.789          21-22-66            -
  32   NDYFCML   33  7.852  121.066  175.945  176.615  53.657  58.412    35.787   26.201                 -           26
  33         V   26  8.037  121.943  175.461  175.969  65.229  53.643    28.392   35.924                33           50
```

correction 0.0 ppm tol: CO 0.3 CA 0.3 CB 0.3 Trust user assign □

User assign vs result: 100% >80% <80% 0%

Penalty	Residue range	Segments (ID numbers)
1.124	7-17	11-44-32-74-52-34-72-66-35-46-21
2.246	18-25	31-39-73-28-3-27-60-55
2.112	27-30	58-12-22-36
2.392	32-43	33-26-50-16-42-37-15-45-65-70-23-40
2.104	46-54	75-59-54-18-10-25-5-49-61
0.869	55-59	30-17-2-19-41
1.608	62-82	24-56-76-7-64-62-51-6-43-69-68-57-71-79-13-1-20-14-9-29-4

Penalty: 2.392 Good

```
Peak ID  Rno  Rtp  User   dCO    dCA    dCB    Predicted Res(i-1)  Predicted Res(i)
  ---     -    E    -
  33     32    N   32    0.00   0.00   0.00   WEQHRC              NDYFCML
  26     33    V   33   -0.02   0.01  -0.14   NDYFCML             V
  50     34    K   34    0.02  -0.06   0.00   V                   MVK
  16     35    A   35    0.00   0.04   0.27   MVK                 A
  42     36    K   36    0.00  -0.09  -0.07   A                   MVK
  37     37    I   37   -0.02  -0.09   0.14   MVK                 I
  15     38    Q   38    0.03  -0.18   0.14   I                   QHEWC
  45     39    D   39    0.00  -0.15  -0.14   WQHCE               DLYFI
  65     40    K   40   -0.01   0.22   0.14   DLYFI               KVMCH
  70     41    E   41   -0.01  -0.02  -0.07   KVMCH               MKHREW
  23     42    G   42   -0.02  -0.04   0.14   MKHREW              G
  40     43    I   43   -0.06  -0.17   0.00   G                   DLYFI
```

Fig. 4 continued

(χ^1, χ^2) using the side-chain rotamer library established by Lovell and colleagues (Lovell et al. 2000) (Fig. 8B).

Application example of KUJIRA for a small protein, human ubiquitin

Uniformly $^{13}C/^{15}N$-labeled human ubiquitin (~0.8 mM), expressed and purified using the methods established by Kigawa and coworkers (Kigawa et al. 1999, 2004), was used for the NMR measurements in $^1H_2O/^2H_2O$ (9:1), containing 20 mM sodium phosphate buffer (pH 6.0), 100 mM NaCl, 1 mM 1,4-DL-dithiothreitol-d_{10} and 0.02% NaN_3. This human ubiquitin construct comprises 83 amino acid residues, with 76 residues of the natural sequence and 7 N-terminal artificial residues, GSSGSSG.

All of the 2D and 3D NMR spectra were measured at 25°C on Bruker AVANCE600 and AVANCE800 spectrometers. The acquired spectral data were processed with

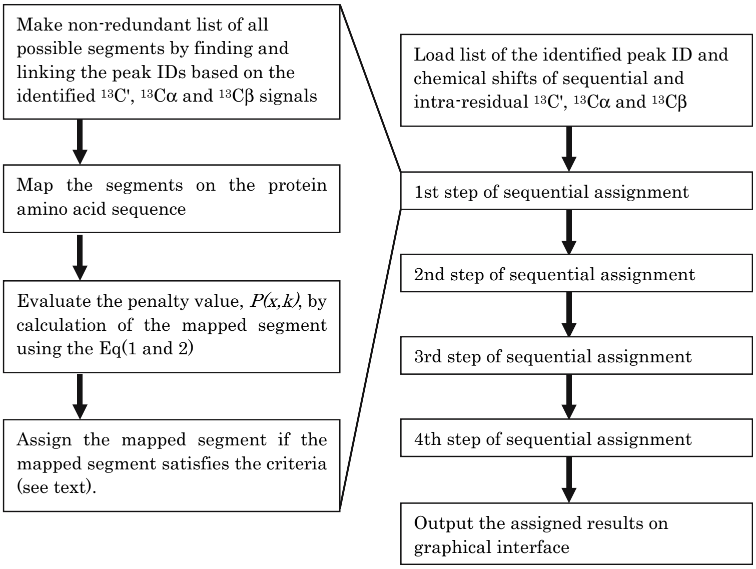

Make non-redundant list of all possible segments by finding and linking the peak IDs based on the identified ^{13}C', ^{13}Cα and ^{13}Cβ signals
Map the segments on the protein amino acid sequence
Evaluate the penalty value, *P(x,k)*, by calculation of the mapped segment using the Eq(1 and 2)
Assign the mapped segment if the mapped segment satisfies the criteria (see text).
Load list of the identified peak ID and chemical shifts of sequential and intra-residual ^{13}C', ^{13}Cα and ^{13}Cβ
1st step of sequential assignment
2nd step of sequential assignment
3rd step of sequential assignment
4th step of sequential assignment
Output the assigned results on graphical interface

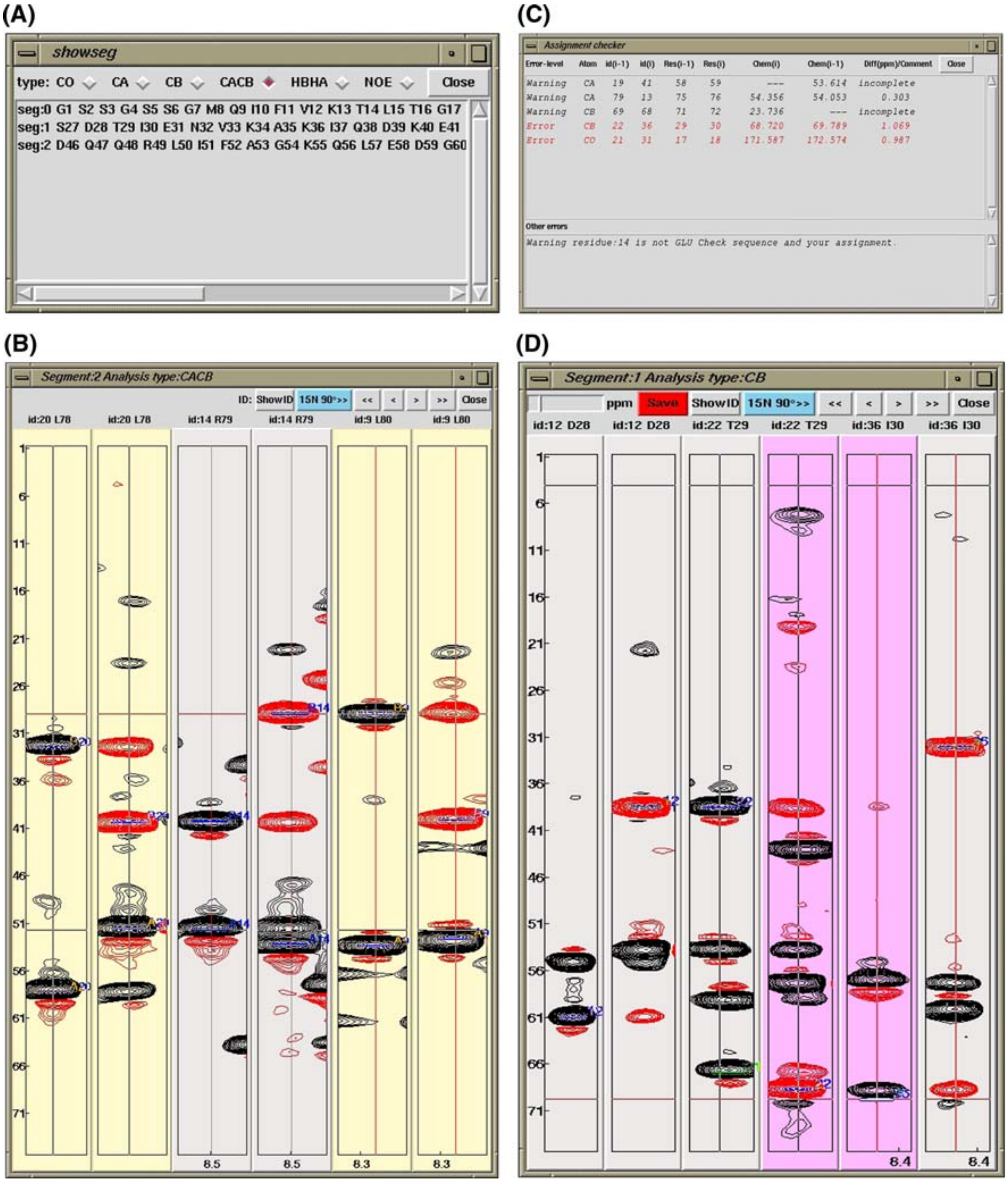
Fig. 5 The flow chart of the automated sequence-specific main-chain assignment, using the program "QuickAssign". The first step of the program job is loading a text file describing the basic data for the assignment, including the chemical shifts of ^{1}H, ^{15}N, sequential and intra-residual $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and $^{13}C'$ signals, for the user-defined peak IDs. The chart on the left describes the flow chart of each sequential assignment step

the program NMRPipe (Delaglio et al. 1995). The 3D triple resonance experiments, HNCO, HN(CA)CO, HN(CO)CA, HNCA, CBCA(CO)NH, and HNCACB, were used for the chemical shifts of the sequence-specific main-chain and $^{13}C_{\beta}$ signal assignments, while HBHA(CBCACO)NH, C(CCO)NH, and HCCH-TOCSY were used for the chemical shift assignments of the side-chain signals. ^{15}N-edited NOESY and ^{13}C-edited NOESY (covering both aliphatic and aromatic region of ^{13}C dimension) experiments with a mixing time of 80 ms were used for the main-chain and side-chain assignments, as well as for obtaining the NOE peaks.

Combined automated NOE cross-peak assignment by the CANDID algorithm (Herrmann et al. 2002) and structure calculation by torsion angle dynamics (Güntert et al. 1997) were performed using the software package CYANA 1.0.7 (Güntert 2003). Peak lists for the ^{15}N-edited NOESY and ^{13}C-edited NOESY spectra were generated by automated and manual peak picking with NMRView. The input data files also included a chemical shift table for the assigned signals, which was managed by KUJIRA. The backbone and C^{β} chemical shift values were analyzed with the program TALOS (Cornilescu et al. 1999) to generate dihedral angle constraints with lower and upper limits of ±30 degrees around the most probable values of the ϕ and ψ backbone dihedral angles. No other structural constraints were applied in the CYANA calculations. Twenty IBM Power4 processors (1.5 GHz) of an IBM p655 server were employed for the CYANA calculations. An accurate solution structure of human ubiquitin was previously determined by Cornilescu and colleagues (Cornilescu et al. 1998), on the basis of the virtually complete chemical shift assignments and a large number of structural constraints, including 2,727 NOE, 98 dihedral angle, 1,307 residual dipolar coupling and 27 hydrogen bond related constraints. The coordinate data (PDB Accession code 1D3Z) and the chemical shift data (BMRB Accession Number 6457) were used as a reference for comparison with the structures and chemical shifts determined in this study.

Results and discussion

As of this year, our research group has determined more than 1,200 protein structures by NMR techniques, as a part of the RIKEN Structural Genomics/Proteomics Initiative (RSGI) for the "Protein 3000" Project in Japan. At present, more than 800 of these structures can be found in the Protein Data Bank, which were solved by means of a protocol similar to the one employing KUJIRA and CYANA as described in the following. The conventional strategy for NMR protein structure determination follows the paradigm that an accurately determined structure is the consequence of accurately assigned NOEs, which, in turn, are derived from accurately determined chemical shifts. Ideally, each stage of the structure analysis, as mentioned in the introduction, should be completed with high accuracy and completeness, before starting the next one. Otherwise, it would be difficult to obtain a correct structure. However, the exacting requirements of the previous stage can delay the start of the subsequent stage, thus slowing down the whole process. This constitutes a weakness of the conventional strategy, if it is applied to high-throughput NMR analysis. Another weakness could arise if the conventional strategy is not facilitated with an integrated platform. Since nowadays, structure analyses by NMR require a number of software packages, the NMR scientist has to be experienced with each of them. The data exchange, such as importing and exporting from one software format to another, is more complicated as the number



type: CO ◇ CA ◇ CB ◇ CACB ◆ HBHA ◇ NOE ◇

```
seg:0 G1 S2 S3 G4 S5 S6 G7 M8 Q9 I10 F11 V12 K13 T14 L15 T16 G17
seg:1 S27 D28 T29 I30 E31 N32 V33 K34 A35 K36 I37 Q38 D39 K40 E41
seg:2 D46 Q47 Q48 R49 L50 I51 F52 A53 G54 K55 Q56 L57 E58 D59 G60
```

Error-level	Atom	id(i-1)	id(i)	Res(i-1)	Res(i)	Chem(i)	Chem(i-1)	Diff(ppm)/Comment
Warning	CA	19	41	58	59	---	53.614	incomplete
Warning	CA	79	13	75	76	54.356	54.053	0.303
Warning	CB	69	68	71	72	23.736	---	incomplete
Error	CB	22	36	29	30	68.720	69.789	1.069
Error	CO	21	31	17	18	171.587	172.574	0.987

```
Warning residue:14 is not GLU Check sequence and your assignment.
```

Fig. 6 (**A**) Pop-up graphical interface showing the list of sequence segments that are expected from the amino acid sequence of the protein. The user can select the type of signals by clicking the checkboxes on the top of the interface. By double-clicking one of the segments, the arrayed 2D-spectrum strips appear (**B**), showing the sequential connectivity of the assigned peak IDs. Each pair of strips corresponds to a peak ID, in which the sequential strip is placed on the left, while the intra-residual strip is on the right. The identified carbon signals are indicated by the blue boxes labeled with the assigned residue number and amino acid type in one-letter code. (**C**) Pop-up graphical interface displaying the results of the function “CheckAssign”. This function examines the sequential connectivities of the assigned peak IDs, and gives a warning message for chemical shift differences of 0.3–0.5 ppm and an error message for those >0.5 ppm. The warning messages with the label “- - -” indicate that the chemical shift is not defined for making the sequential connectivity. The lower list-box displays the warning messages if a redundancy is found in the assigned residue number or the defined peak ID number, or if the amino acid type of the assigned peak ID or the combination of the identified $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts does not match with the actual protein sequence. (**D**) By double-clicking one of the warning/error messages on the upper list-box, a pop-up interface appears to display the arrayed 2D spectrum strips. The problematic sequential connectivity is highlighted by the magenta coloring of the two corresponding strips. The user is allowed to correct and save the identified peak position by directly clicking on the spectrum strip

Table 1 The results of the four steps of automated sequential assignment performed by the external C program “QuickAssign”[a,b]

Penalty	Residue	Assigned segments
First step		*Tol*[c]: 0.3, 0.3, 0.3, *Amb*[d]: 0, *Maxp*[e]: 2.4, *Minc*[f]: 2, *Minl*[g]: 8
1.124	7–17	11-44-32-74-52-34-72-66-35-46-21
2.392	32–43	33-26-50-16-42-37-15-45-65-70-23-40
2.104	46–54	75-59-54-18-10-25-5-49-61
1.608	62–82	24-56-76-7-64-62-51-6-43-69-68-57-71-79-13-1-20-14-9-29-4
Completeness of assignment[h]: 67.1%		
Second step		*Tol*[c]: 0.3, 0.3, 0.3, *Amb*[d]: 1, *Maxp*[e]: 4.0, *Minc*[f]: 2, *Minl*[g]: 8
1.124	7–17	11-44-32-74-52-34-72-66-35-46-21
2.246	18–25	31-39-73-28-3-27-60-55
2.392	32–43	33-26-50-16-42-37-15-45-65-70-23-40
2.104	46–54	75-59-54-18-10-25-5-49-61
1.608	62–82	24-56-76-7-64-62-51-6-43-69-68-57-71-79-13-1-20-14-9-29-4
Completeness of assignment[h]: 77.2%		
Third step		*Tol*[c]: 0.3, 0.3, 0.3, *Amb*[d]: 2, *Maxp*[e]: 4.0, *Minc*[f]: 2, *Minl*[g]: 5
1.124	7–17	11-44-32-74-52-34-72-66-35-46-21
2.246	18–25	31-39-73-28-3-27-60-55
0.869	55–59	30-17-2-19-41
2.392	32–43	33-26-50-16-42-37-15-45-65-70-23-40
2.104	46–54	75-59-54-18-10-25-5-49-61
1.608	62–82	24-56-76-7-64-62-51-6-43-69-68-57-71-79-13-1-20-14-9-29-4
Completeness of assignment[h]: 83.5%		
Fourth step		*Tol*[c]: 0.3, 0.3, 0.3, *Amb*[d]: –, *Maxp*[e]: 6.0, *Minc*[f]: 2, *Minl*[g]: 3
1.124	7–17	11-44-32-74-52-34-72-66-35-46-21
2.246	18–25	31-39-73-28-3-27-60-55
2.112	27–30	58-12-22-36
0.869	55–59	30-17-2-19-41
2.392	32–43	33-26-50-16-42-37-15-45-65-70-23-40
2.104	46–54	75-59-54-18-10-25-5-49-61
1.608	62–82	24-56-76-7-64-62-51-6-43-69-68-57-71-79-13-1-20-14-9-29-4
Completeness of assignment[h]: 88.6%		

[a] The experiment was carried out for the tutorial sample, as described in the text

[b] The basic main-chain assignment information, including peak IDs, the chemical shifts of $^{1}H_{N}$, $^{15}N\alpha$, sequential and intra-residual $^{13}C'$, $^{13}C\alpha$ and $^{13}C\beta$, was prepared manually

[c] Tolerance values (ppm) for identifying the sequential connectivity of the peak ID, based on ^{13}C signals respectively corresponding to $^{13}C'$, $^{13}C\alpha$ and $^{13}C\beta$

[d] Maximal ambiguity of the sequential connectivity

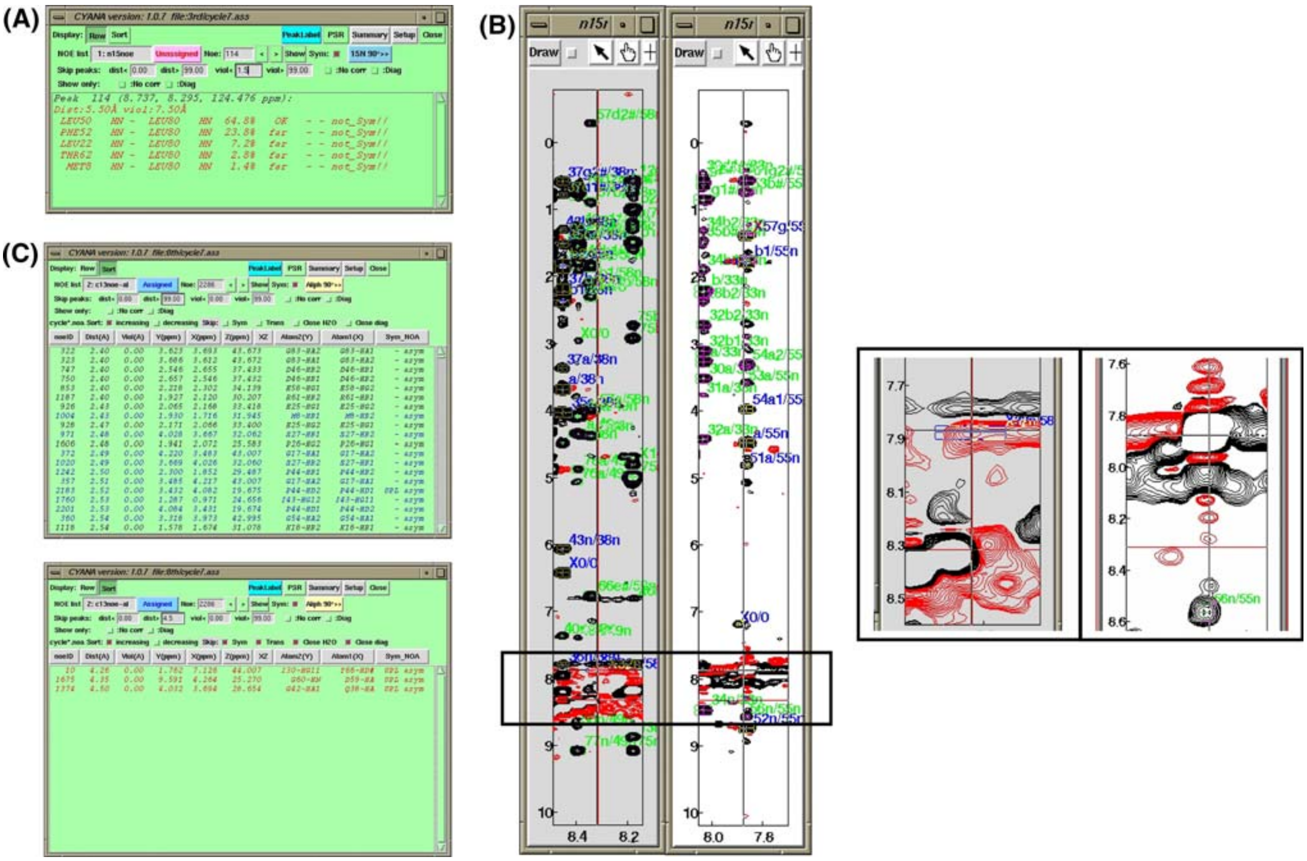
[e] Maximal penalty value to assign the mapped segment

[f] Minimal number of signals used for identifying the sequential connectivity

[g] Minimal length of segment to be judged

[h] The total number of residues to be assigned is 78

of required software programs increases. As mentioned above, the CCPN and SPINS data models are a promising solution to address these problems, by developing a program suite to integrate many software programs, NMR data and related information. The remarkable feature of the CCPN data model is not only the reduction in the tedious manipulation of NMR data but also their originally created API, with which scientists within the community can easily develop software and maintain it for long time by collaborations with others in the community. In contrast to the data models, KUJIRA is designed to be a package of modules that are as small as possible, by restricting the functions of each module to those essentially required for the analysis. KUJIRA is highly optimized for high-throughput NMR studies by the sophisticated subroutines working on Tcl/Tk and the external C programs utilized for



```
CYANA version: 1.0.7  file:3rd/cycle7.ass
Display: Row  Sort                                   PeakLabel  PSR  Summary  Setup  Close
NOE list  1: n15noe  Unassigned  Noe: 114  <  >  Show  Sym: ■  15N 90°>>
Skip peaks:  dist< 0.00  dist> 99.00  viol< 1.5  viol> 99.00  □ :No corr  □ :Diag
Show only:  □ :No corr  □ :Diag
Peak  114 (8.737, 8.295, 124.476 ppm):
Dist:5.50Å viol:7.50Å
 LEU50   HN -  LEU80   HN  64.8%   OK   - - not_Sym!!
 PHE52   HN -  LEU80   HN  23.8%  far   - - not_Sym!!
 LEU22   HN -  LEU80   HN   7.2%  far   - - not_Sym!!
 THR62   HN -  LEU80   HN   2.8%  far   - - not_Sym!!
  MET5   HN -  LEU80   HN   1.4%  far   - - not_Sym!!
```

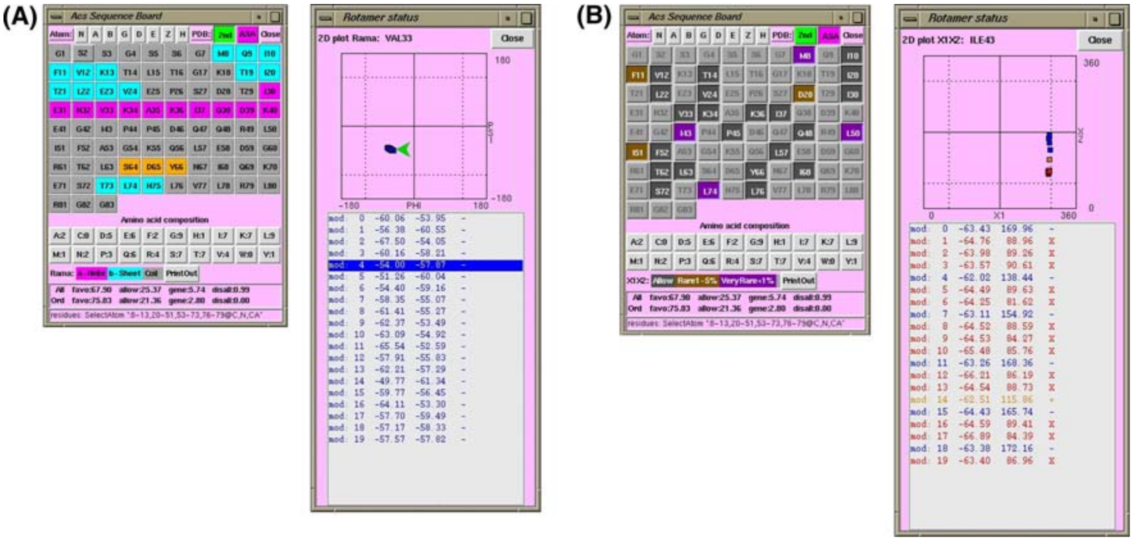
```
CYANA version: 1.0.7  file:8th/cycle7.ass
Display: Row  Sort                                   PeakLabel  PSR  Summary  Setup  Close
NOE list  2: c13noe-al  Assigned  Noe: 2286  <  >  Show  Sym: ■  Aliph 90°>>
Skip peaks:  dist< 0.00  dist> 99.00  viol< 0.00  viol> 99.00  □ :No corr  □ :Diag
Show only:  □ :No corr  □ :Diag
cycle*.noa Sort: ■ increasing  □ decreasing  Skip: □ Sym  □ Trans  □ Close H2O  □ Close diag
noeID  Dist(A)  Viol(A)  Y(ppm)  X(ppm)  Z(ppm)  XZ  Atom2(Y)  Atom1(X)  Sym_NOA
  322   2.40   0.00   3.623   3.693   43.673    G83-HA2   G83-HA1    - asym
  323   2.40   0.00   3.606   3.612   43.672    G83-HA1   G83-HA2    - asym
  747   2.40   0.00   2.546   2.655   37.433    D46-HB2   D46-HB1    - asym
  750   2.40   0.00   2.657   2.546   37.432    D46-HB1   D46-HB2    - asym
  853   2.40   0.00   2.218   2.302   34.139    E58-HG1   E58-HG2    - asym
 1187   2.40   0.00   1.927   2.120   30.207    R61-HB2   R61-HB1    - asym
  926   2.43   0.00   2.065   2.168   33.418    E25-HG1   E25-HG2    - asym
 1004   2.43   0.00   1.930   1.716   31.945     M8-HB1    M8-HB2    - asym
  928   2.47   0.00   2.171   2.066   33.400    E25-HG2   E25-HG1    - asym
  971   2.48   0.00   4.028   3.667   32.062    S27-HB1   S27-HB2    - asym
 1606   2.48   0.00   1.941   2.072   25.583    P26-HG2   P26-HG1    - asym
  372   2.49   0.00   4.220   3.483   43.007    G17-HA1   G17-HA2    - asym
 1020   2.49   0.00   3.669   4.026   32.060    S27-HB2   S27-HB1    - asym
 1242   2.50   0.00   2.300   1.852   29.487    P44-HB1   P44-HB2    - asym
  357   2.51   0.00   3.485   4.217   43.007    G17-HA2   G17-HA1    - asym
 2183   2.52   0.00   3.432   4.082   19.675    P44-HD2   P44-HD1  SYL asym
 1760   2.53   0.00   1.287   0.971   24.656   I43-HG12  I43-HG11    - asym
 2201   2.53   0.00   4.084   3.431   19.674    P44-HD1   P44-HD2    - asym
  360   2.54   0.00   3.318   3.973   42.995    G54-HA2   G54-HA1    - asym
 1118   2.54   0.00   1.578   1.674   31.078    K16-HB2   K16-HB1    - asym
```

```
CYANA version: 1.0.7  file:8th/cycle7.ass
Display: Row  Sort                                   PeakLabel  PSR  Summary  Setup  Close
NOE list  2: c13noe-al  Assigned  Noe: 2286  <  >  Show  Sym: ■  Aliph 90°>>
Skip peaks:  dist< 0.00  dist> 4.5  viol< 0.00  viol> 99.00  □ :No corr  □ :Diag
Show only:  □ :No corr  □ :Diag
cycle*.noa Sort: ■ increasing  □ decreasing  Skip: ■ Sym  ■ Trans  ■ Close H2O  ■ Close diag
noeID  Dist(A)  Viol(A)  Y(ppm)  X(ppm)  Z(ppm)  XZ  Atom2(Y)  Atom1(X)  Sym_NOA
   10   4.26   0.00   1.762   7.126   44.007   I30-HG11   F66-HD#  SYL asym
 1675   4.35   0.00   9.591   4.264   25.270     G60-HN    D59-HA  SYL asym
 1374   4.50   0.00   4.032   3.694   28.654    G42-HA1    Q38-HA  SYL asym
```

Fig. 7 (**A**) Graphical interface of the CYANA analysis module of KUJIRA, for the evaluation of fully automated NOE assignments carried out by CYANA. The module can load the exported log file for the CYANA automated NOE assignment calculation. The list-box represents the NOE assignment information from the CYANA log file for a certain peak eliminated by a large distance violation. The NOE assignment candidates are ordered by their generalized volume contribution, as determined by the automated NOE assignment algorithm CANDID (Herrmann et al. 2002) in CYANA. The buttons in the upper part of the interface are used to increase or decrease the peak ID number, to switch between the lists of unassigned and assigned NOE peaks, and to flip the displayed 2D spectrum strips by 90 degrees about the *z*-axis. The "Skip peaks" and "Show only" items serve to filter the peaks for which information can be displayed according to various criteria. (**B**) An example display of a typical artifact peak that has been eliminated for the fully automated NOE assignment. The 2D spectrum strips brought up by the Sync-Jump function of the CYANA analysis module show the narrow region of the ^{15}N edited NOESY spectrum, corresponding to the Lys55 H_N-Nα-signal, alongside the spectrum strip showing the transposed position of the NOE peaks according to the assignment, Lys55 H_N–Glu58 H_N. The right panel represents expanded portions of the two 2D spectrum strips, showing the unassigned artifact peak at the position (7.9, 8.3, 123.1 ppm), while no peak is found at the transposed position of the NOE assignment (8.3, 7.9, 121.7 ppm). Using the CYANA analysis module with the skip function, the user can quickly access this peak and determine whether it is an artifact. (**C**) The graphical interface of the CYANA analysis module in the "Sort" mode, representing NOE peaks sorted, for instance, by their upper distance limit derived from the peak intensity. In the upper panel, all 2,262 assigned peaks derived from 3D ^{13}C edited NOESY are listed without any skip setting (only the top 20 peaks appear, because of the limited height of the list-box). The lower panel represents the subset of the NOE peaks obtained by skipping the peaks with an estimated ^{1}H–^{1}H distance bound longer than 4.5 Å, or located closer than 0.1 ppm from the water line or the diagonal, or having a symmetrical consensus for the prime candidate of the NOE assignment. Only three NOE peaks selectively appear in the list. Strong NOE peaks lacking the symmetrical consensus of the NOE assignment would be very useful indicators of potential errors in the chemical shift assignments

the heavy tasks, which would not run fast in the interpreter based programming language. Owing to the compactness of the package, the installation of KUJIRA has been simplified and does not require any specific library or third party software, except for NMRView and Tcl/Tk, which are also simple and easy to install.

Based on our experience with a large number of NMR structure studies, we have established a robust and expeditious strategy for NMR protein structure determination using KUJIRA and CYANA, which represents an innovation over the conventional methods. The strategy, shown schematically in Fig. 9, is simplified into three analysis phases; namely, the early phase, "Preliminary assignment and structure analysis", the intermediate phase, "Refinement of assignment and structure analysis" and the final phase, "Finalization of


(A)
Acs Sequence Board
Atom: N A B G D E Z H PDB: 2nd ASA Close
G1 S2 S3 G4 S5 S6 G7 M8 Q9 I10
F11 V12 K13 T14 L15 T16 G17 K18 T19 I20
T21 L22 E23 V24 E25 P26 S27 D28 T29 I30
E31 N32 V33 K34 A35 K36 I37 Q38 D39 K40
E41 G42 I43 P44 P45 D46 Q47 Q48 R49 L50
I51 F52 A53 G54 K55 Q56 L57 E58 D59 G60
R61 T62 L63 S64 D65 Y66 N67 I68 Q69 K70
E71 S72 T73 L74 H75 L76 V77 L78 R79 L80
R81 G82 G83
Amino acid composition
A:2 C:0 D:5 E:6 F:2 G:9 H:1 I:7 K:7 L:9
M:1 N:2 P:3 Q:6 R:4 S:7 T:7 V:4 W:0 Y:1
Rama: a-Helix b-Sheet Coil PrintOut
Rotamer status
2D plot Rama: VAL33
Close
180
-180
-180
PHI
180
PSI
(B)
Acs Sequence Board
Atom: N A B G D E Z H PDB: 2nd ASA Close
X1X2: Allow Rare1-5% VeryRare<1% PrintOut
Rotamer status
2D plot X1X2: ILE43
Close
360
0
0
X1
360
X2

```
All  favo:67.90  allow:25.37  gene:5.74  disall:0.99
Ord  favo:75.83  allow:21.36  gene:2.80  disall:0.00
residues: SelectAtom '8-13,20-51,53-73,76-79@C,N,CA'
```

```
mod:  0  -60.06  -53.95   -
mod:  1  -56.38  -60.55   -
mod:  2  -67.50  -54.05   -
mod:  3  -60.16  -58.21   -
mod:  4  -54.00  -57.87   -
mod:  5  -51.26  -60.04   -
mod:  6  -54.40  -59.16   -
mod:  7  -58.35  -55.07   -
mod:  8  -61.41  -55.27   -
mod:  9  -62.37  -53.49   -
mod: 10  -63.09  -54.92   -
mod: 11  -65.54  -52.59   -
mod: 12  -57.91  -55.83   -
mod: 13  -62.21  -57.29   -
mod: 14  -49.77  -61.34   -
mod: 15  -59.77  -56.45   -
mod: 16  -64.11  -53.30   -
mod: 17  -57.70  -59.49   -
mod: 18  -57.17  -58.33   -
mod: 19  -57.57  -57.82   -
```

```
All  favo:67.90  allow:25.37  gene:5.74  disall:0.99
Ord  favo:75.83  allow:21.36  gene:2.80  disall:0.00
residues: SelectAtom '8-13,20-51,53-73,76-79@C,N,CA'
```

```
mod:  0  -63.43  169.96   -
mod:  1  -64.76   88.96   X
mod:  2  -63.98   89.26   X
mod:  3  -63.57   90.61   X
mod:  4  -62.02  138.44   -
mod:  5  -64.49   89.63   X
mod:  6  -64.25   81.62   X
mod:  7  -63.11  154.92   -
mod:  8  -64.52   88.59   X
mod:  9  -64.53   84.27   X
mod: 10  -65.48   85.76   X
mod: 11  -63.26  168.36   -
mod: 12  -66.21   86.19   X
mod: 13  -64.54   88.73   X
mod: 14  -62.51  115.86   +
mod: 15  -64.43  165.74   -
mod: 16  -64.59   89.41   X
mod: 17  -66.89   84.39   X
mod: 18  -63.38  172.16   -
mod: 19  -63.40   86.96   X
```

Fig. 8 (**A**) Graphical interface of the structure validation module in the secondary structure display mode (left panel). The module calculates the hydrogen-bond based secondary structure assignment in the specified NMR structures. The detected secondary structures are indicated by different colors: cyan: β-strand, magenta: α-helix, and orange: 3_{10}-helix. In the pop-up graphical interface on the right, the 2D Ramachandran plot analysis and the ϕ and ψ dihedral angles of residue Val33 in 20 structure models are represented. (**B**) The graphical interface of the structure validation module in the side-chain structure analysis mode (left panel). The module calculates the average solvent accessible surface area for the side-chain of each residue in the specified NMR structures. The residues with solvent accessible surface area values less than 30% are classified as being involved in the core of the protein, and are represented by sunken buttons. Problematic χ^1/χ^2 dihedral angle pairs are indicated in different colors, based on their probability, as analyzed using a standard rotamer library; dark orange indicates rare (probability 1–5%), and dark magenta indicates very rare (less than 1%) conformations. In the pop-up graphical interface on the right, the 2D χ^1/χ^2 plot analysis and the χ^1 and χ^2 dihedral angle values of Ile43 in 20 structure models are represented

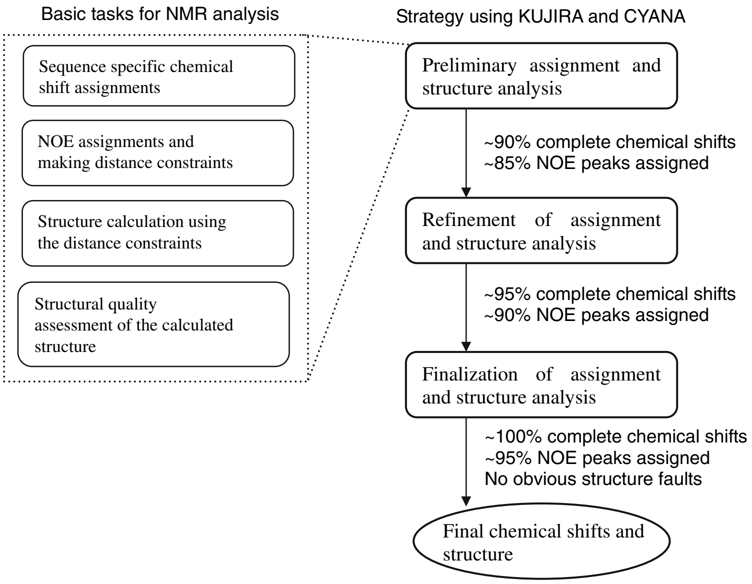

Basic tasks for NMR analysis
Sequence specific chemical shift assignments
NOE assignments and making distance constraints
Structure calculation using the distance constraints
Structural quality assessment of the calculated structure
Strategy using KUJIRA and CYANA
Preliminary assignment and structure analysis
~90% complete chemical shifts
~85% NOE peaks assigned
Refinement of assignment and structure analysis
~95% complete chemical shifts
~90% NOE peaks assigned
Finalization of assignment and structure analysis
~100% complete chemical shifts
~95% NOE peaks assigned
No obvious structure faults
Final chemical shifts and structure

Fig. 9 High-throughput strategy for NMR structure determinations based predominantly on NOE-based structural constraints, using KUJIRA and CYANA. Every stage includes all of the basic tasks for NMR analysis: interactive assessment and correction of chemical shift and NOE assignments and structure calculation. The guidelines are also indicated on the right side of the flow chart. The typical methods to collect NOE peaks for CYANA calculations are mentioned in the section "Materials and methods" describing the methods for the application example of ubiquitin. To evaluate the completeness of the chemical shift assignments, all of the main-chain and side-chain ^{1}H, ^{13}C, and ^{15}N signals responsible for constructing the defined region of protein, such as main-chain, buried aliphatic and aromatic side-chain signals, should be involved. The signals that cannot be observed in the 3D NOESY because of chemical exchange or bad water suppression should be also considered to be "missing" signals. For the CYANA calculation in any stage of NMR analysis, all of the observed NOE peaks should be applied. To pass the finalization stage, the NOEs that could not be assigned by CYANA should not be clustered in a certain region of the calculated structure

assignment and structure analysis". The most remarkable feature of the strategy is that the chemical shift assignments and the structure determination cooperatively progress toward the final state of the analysis, rather than reaching the end result through a sequence of strictly consecutive steps.

Preliminary assignment and structure analysis

In the early phase of the analysis, referred to as the "Preliminary assignment and structure analysis," the chemical shift assignments are nearly completed with KUJIRA prior to the first structure calculation by CYANA. The most crucial step in this analysis stage must be the sequence-specific main-chain signal assignments, since the following side-chain signal assignments and NOE peak assignments strongly rely on the assignments. Many automated sequence-specific assignment programs have been reported, based on a variety of assignment algorithms, such as Monte Carlo algorithms (Buchler et al. 1997; Leutner et al. 1998; Lukin et al. 1997), genetic algorithms (Bartels et al. 1996a, 1996b), exhaustive search algorithms (Andrec and Levy 2002; Atreya et al. 2000; Coggins and Zhou 2003; Güntert et al. 2000), heuristic comparisons to predicted shifts (Gronwald et al. 1998), and heuristic best-first algorithms (Hyberts and Wagner 2003; Li and Sanctuary 1997; Zimmerman et al. 1994; Zimmerman et al. 1997). The difficulties found with the automated assignment programs include several missing signals in 2D or 3D spectra, which would be encountered in the assignments of any NMR samples even those where the ^{1}H–^{15}N HSQC spectra show well-separated signals. The existence of minor signals, which might be derived from exchanges between major/minor conformations, and some contamination would also be obstacles for the assignment jobs. One of the biggest advantages of the semi-automated main-chain signal assignment modules in KUJIRA is that the program QuickAssign runs very fast and provides reasonably accurate results. As shown in Table 1, the four stages of the iterative assignment jobs were finished within 1 sec on the SGI workstation (MIPS14000, 500 MHz), and no assignment error was found. This high-performance can help the user to accelerate the interactive work on the graphical interfaces, and thus to find and correct the problems with the assignments readily. Typically, in our projects, an assignment time of 1–3 h was required for the complete the assignments of main-chain and ^{1}H$_{\alpha}$ and ^{1}H$_{\beta}$ signals using KUJIRA. The remaining side-chain signals were manually assigned, using the chemical shift database module of KUJIRA, to achieve more than 90% of ^{1}H, ^{15}N, and ^{13}C signals. This process typically takes for 5–8 h, if HCCH-TOCSY and ^{13}C-edited NOESY spectra with reasonable quality are available.

After automated peak picking of the 3D NOESY spectra and filtering of the noise peaks, the first CYANA calculation is performed. The major purpose of the CYANA calculation(s) in this phase is to obtain tentative structures with correct chain-folds and core-packing and to eliminate obviously erroneous NOE peaks. Displaying peak boxes labeled with NOE assignment information on the 3D NOESY spectra streamlines this task. Using the CYANA analysis module of KUJIRA, which can incorporate the results of the automated NOE assignment by CYANA and provides the Sync-Jump commands based on the assignments, the user can readily access the problematic peaks and quickly determine whether they are artifacts.

Refinement of assignment and structure analysis

In the middle phase of the analysis, referred to as the "Refinement of assignment and structure analysis," entries in the chemical shift table are corrected, based on the spectrum positions of the NOE peaks without assignments. The values in the chemical shift table that deviate slightly from the positions of a cluster of NOE peaks may suggest that a minor error exists in the table. Such a minor error might not be serious enough to disrupt the global fold of the calculated structure; however, because of the error, CYANA may fail to assign the indirect dimension of the NOE peak, if the related atoms are close together in the calculated structure. This may result in a local fault in the calculated structure, and therefore, it should be corrected in this analysis stage. In particular, a group of unassigned NOE peaks that are apparently clustered and aligned on a certain *x*–*z* axis position, as illustrated in Fig. 10A, is often indicative of a slight error in the chemical shift table. For instance, based on such aligned, unassigned peaks, a methylene group that was mistakenly assumed to yield degenerate signals can effectively be recognized as separate methylene signals. Owing to the robust NOE assignment capability of CYANA, many of the spurious NOEs are successfully removed at the end of the middle phase of the analysis. To proceed to the final stage of the analysis, we nevertheless have to consider the possibility that a very small number of artifact peaks and some missing signals might have evaded the manual analysis. These artifact peaks would more seriously distort a loosely defined structure, such as surface, loop and turn regions, rather than a well-defined region. Therefore, the refinement work in the middle phase of the structure analysis aims at identifying the few remaining artifact peaks and the inaccuracies of a small number of chemical shifts among the thousands in the NMR data sets. In most cases, the remaining artifact peaks tend to be related to signals that are both nearly degenerate in the 2D-HSQC spectrum

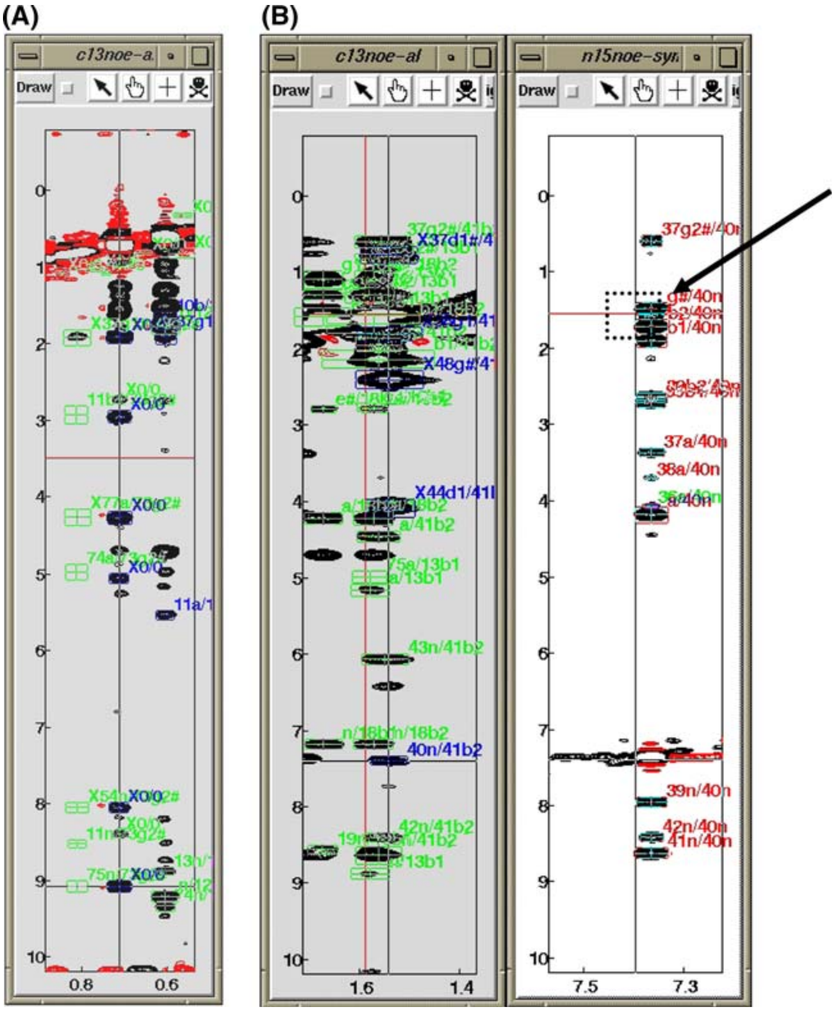


Fig. 10 (**A**) Example of a group of aligned peaks, derived from a 3D ^{13}C edited NOESY, which were unassigned in the fully automated NOE assignment carried out by CYANA. The labels "X0/0" indicate peaks that were unassigned because no corresponding signal was found for the direct and/or indirect dimensions in the NOE assignment. The 2D spectrum strip shows the region close to Val77-Hγ1, where the clustered peaks are located slightly away from the position specified in the chemical shift table. Using the CYANA analysis module applied with the skip status "show only no corresponding signal," the user can easily and quickly access a group of unassigned NOE peaks. In combination with the peak labeling function, the user can straightforwardly notice inaccuracies in the chemical shift assignment data. (**B**) A NOE assignment lacking "symmetry." A NOE peak in the 2D strip corresponding to the signal of Glu41-Hβ2, derived from 3D ^{13}C edited NOESY, has been assigned to Lys40-HN (lower part of the left panel indicated by the label "40n/41b2"). No NOE peak was found on the transposed position of the NOE assignment (pointed out by the arrow and the dotted box in the right panel); Glu41-Hβ2 on the 2D spectrum strip of Lys40-H$_N$ in the 3D ^{15}N-edited NOESY spectrum. The erroneous NOE assignment was caused by mistakenly assuming the degeneracy of the two γ-methylene proton signals of Gln48 during the chemical shift assignment. Using the CYANA analysis module of KUJIRA, the problem was noticed in the 7th stage of the tutorial analysis, and was fixed in the Chemical shift database module. See text for details

projection, and originating from atoms that are spatially proximate in the protein structure. Such peaks might not cause a large violation of the NOE assignment, which is based on both the chemical shifts and inter-proton distances, and may therefore lead CYANA to apply the corresponding distance constraint in the structure calculations. The work required to identify these artifact peaks among the few thousand peaks can be awfully tedious, even if the user carefully operates the spectrum windows, unless some mechanisms help the user to intuitively notice them. The most remarkable aspect of the CYANA analysis module in KUJIRA is the function to evaluate the symmetrical consensus of the NOE assignment, by detection of the peak intensity at the transposed position deduced from the NOE assignment. This module highlights the assignments of potentially spurious NOE peaks, if the spectral intensity at the transposed position is below a user-specified threshold (see Fig. 7C).

Finalization of assignment and structure analysis

In the final phase of the analysis, referred to as the "Finalization of assignment and structure analysis," the Structure quality assessment module is used to pinpoint the possible existence of remaining problems in the chemical shift assignments and the NOE peak table. The Structure quality assessment module calculates the secondary structure, the solvent accessible surface area of the side-chain groups, and the order parameter of the dihedral angles. Based on the structural information, the user can readily infer which residues are responsible for constructing the protein structure. The module also has a function to visualize the Ramachandran plot for the ϕ and ψ angles, and an analogous plot for the χ^1 and χ^2 angles. By reference to standardized rotamer libraries, the user can identify the residues that have a conformational problem. Figure 10B shows a NOE peak, found in the NOE peak table, that apparently lacks the symmetrical consensus of its NOE assignment, as detected by the CYANA analysis module, and that generated a spurious distance constraint. It would be very time-consuming to find this kind of problem hidden in the enormous amount of NMR data by the conventional methods; however, the combination of the above-mentioned strategies can make it possible, in a systematic and efficient manner.

Quick and accurate structure determination accomplished by a beginner

As an application example for the case of NMR analysis, the solution structure of human ubiquitin was determined by a post-graduate student who lacked experience in protein structure determination, but was well educated in biochemistry and organic chemistry. The analysis was a tutorial for learning structure analysis, according to the new strategy using KUJIRA and CYANA, under the supervision of an expert NMR scientist. The details of the analysis, summarized in Table 2, comprised nine stages: stages 1–3, corresponding to "Preliminary assignment and structure analysis", stages 4–6, "Refinement of assignment and structure analysis", and stages 7–9, "Finalization of assignment and structure analysis." The semi-automatic assignments for the main-chain and some side-chain signals, including $^1H_\alpha$ and $^1H_\beta$ protons, were assigned by means of the described methods in approximately 3 h. The remaining side-chain signal assignments were finished in 5 h, and the first CYANA calculation was performed (stage 1). At this stage, more than 90% of the ^{1}H, ^{15}N, and ^{13}C signals that could be observed in the measured 2D and 3D spectra were assigned. Another 4 days were spent for eight stages of CYANA calculations, to further refine the chemical shift assignments and the NOE peak table and structure quality assessments, using the KUJIRA modules. After the tutorial analysis work, the accuracy of the assigned chemical shifts and the determined structure were assessed by comparison with the reference structure (1D3Z from PDB) and the reference chemical shifts (bmrb6457.str from BMRB). The accuracy of the assigned chemical shifts was examined by counting the number of incorrect values, with errors from the reference data greater than 0.1, 1.0, and 1.0 ppm for the ^{1}H, ^{13}C, and ^{15}N signals, respectively. Except for the signals of the two N-terminal residues and a His residue, for which the chemical shifts might be affected by slight differences in the protein construct and the experimental conditions, more than 97% of the chemical shifts (779 out of 796) determined by the beginner nearly matched those in the reference data. A remarkable result, shown in Table 2, is that the structure of the first CYANA calculation already yielded the correct chain-fold and core-packing, as compared with the reference solution structure of human ubiquitin, showing 1.02 Å and 1.20 Å RMSD values for the main-chain and side-chain structures, respectively. All χ^1 angles of the residues involved in the hydrophobic core were also correct, from the second stage onwards. These facts strongly support the high reliability of CYANA to calculate the correct structure, in spite of the initially incomplete NMR data. The additional structural constraints, the 86 dihedral angle constraints derived from the TALOS analysis (applied in stages 8–9) and the 10 stereo-specific assignments for nine residues (applied in stages 6–9), did not significantly improve the accuracy of the calculated structures. Alternatively, the chemical shifts of five signals were corrected, and more than 100 erroneous NOE peaks were removed, using the CYANA analysis module in KUJIRA in stages 6–9. In the final CYANA calculation (stage 9), the completeness of the NOE assignments reached more than 97%, involving no NOE peak eliminated by a violation greater than 1.5 Å. No significant deviations from the reference assignments were found in the chemical shift assignments after stage 7. As compared with the reference structure, the structures obtained in the final stage exhibit high accuracy, with low RMSD values for the main-chain (0.66 Å) and side-chain (0.80 Å) atoms. Remarkably, through all of the stages, the increase in the accuracy for the chemical shift assignments agreed well with the increase in the completeness for the automated NOE assignments. This clearly indicates that the high completeness and accuracy of the chemical shift assignments are related to the completeness of the NOE assignment, and supports the accuracy of the assigned chemical shifts, which are responsible for the structure determination. It should be emphasized that the analyst was allowed to concentrate on monitoring the completeness of the chemical shifts and NOE peak tables and the structural faults found in the calculated structures, owing to the fact

Table 2 Summary of structure analysis stages

Stages[a]	Parameters for structure calculations performed by CYANA			NOE assignments		Incorrect assignments in the chemical shift values of ^{1}H, ^{15}N, and ^{13}C signals as compared with those of the 9th stage[f]		Deviation from reference structure			
	Final target function (Å^{2})	Main-chain RMSD[b] (Å)	Violation[c] >0.2 Å	NOE peaks[d]	Completeness[e] (%)	>0.03 ppm for ^{1}H, >0.3 ppm for ^{13}C and ^{15}N	>0.06 ppm for ^{1}H, >0.6 ppm for ^{13}C and ^{15}N	Main-chain RMSD[g] (Å)	Side-chain RMSD[g] (Å)	χ^{1}[h] deviation >60deg	χ^{2}[h] deviation >60deg
Preliminary assignment and structure analysis											
1	11.96	0.34	27	4917	82.3	55	44	1.02	1.20	12(1)	4(3)
2	0.71	0.37	0	3334	89.1	32	28	0.90	1.04	3(0)	4(0)
3	0.79	0.33	0	3203	94.0	30	26	1.00	1.11	3(0)	2(2)
Refinement of assignment and structure analysis											
4	0.84	0.31	1	3274	92.5	13	10	0.93	1.02	4(0)	2(1)
5	0.84	0.28	0	3229	94.8	5	4	0.78	0.88	4(0)	2(2)
6	0.31	0.27	0	3281	93.0	5	4	0.81	0.91	3(0)	2(1)
Finalization of assignment and structure analysis											
7	0.44	0.29	0	3167	96.8	3	2	0.83	0.92	3(0)	1(2)
8	0.68	0.30	0	3164	96.9	0	0	0.79	0.89	3(0)	3(1)
9	0.55	0.29	0	3157	97.1	–	–	0.66	0.80	2(0)	3(1)

Parameters related to the calculated structures are based on the last (7th) CYANA calculation cycle of each stage

[a] In the stages 1–3, 50 structures were calculated in each CYANA calculation cycle, and the 10 structures with the lowest target functions were selected for the subsequent calculation cycles, while in stages 4–9, 100 structures were calculated and 20 structures were selected. ϕ and ψ dihedral angle constraints derived from TALOS predictions were applied in stages 8–9. Stereo-specific assignments were applied in stages 6–9 for the β-methylene groups of Met8, Gln9, Phe11, Leu22 and Glu25, the γ-methylene groups of Gln9 and Ile10, and the δ methyl groups of Leu57, Leu63, and Leu74

[b] RMSD values to the mean coordinates calculated for the backbone atoms, N, C$^{\alpha}$, and C′ of residues 8–13, 20–51, 53–73, and 76–79 after superposition of the calculated structures

[c] Number of used distance constraints with average violation greater than 0.2 Å

[d] Number of cross peaks from the 3D ^{15}N- and ^{13}C-edited NOESY spectra that were used as input for the automated NOE assignment with CYANA

[e] Percentage of NOESY cross peaks assigned by CYANA relative to the total number of NOESY cross peaks

[f] Number of chemical shift assignments that deviate from those of the final (9th) stage by more than the given cutoffs

[g] Deviation of the calculated CYANA structure from the reference structure (1D3Z). RMSD values are between the mean coordinates of the main-chain atoms N, C$^{\alpha}$, and C′, and main-chain and short range side-chain atoms (C′, N, Cα, Cβ and Cγ) of residues 8–13, 20–51, 53–73, and 76–79, after superposition of the calculated structures

[h] Number of residues with χ^{1} or χ^{2} angles deviating more than 60 degrees from the average value in the reference structure. Numbers in parentheses are for the residues in the hydrophobic core of the protein with less than 30% solvent accessible surface area

that the completeness of the chemical shift assignments and the accuracy of the calculated structure were less effective for the subsequent analysis stage. This is the biggest advantage of this new strategy, which makes high-throughput NMR structure studies possible.

Signal assignment and structure validations after the final stage of the structure analysis

The validations of the assigned chemical shifts and the calculated structures have been the most important challenge for the NMR scientist in the last a few decades (Spronk et al. 2004). Several methods are commonly used for the validation; however, there is no universal method to integrate the validation protocols in a small package. Throughout the structural analyses of more than 1,000 protein samples by our research group, we have not seen any significant error in the chain-fold in the determined solution structure of a small (smaller than 20 kDa), stable and monomeric protein that yielded good spectrum quality by our established analysis strategy. It would be nearly impossible to perform an analysis that satisfies our criteria well (see Fig. 9), and then end with a greatly incorrect structure. This is mainly due to the strategy, which is empowered by the intensive corrections of the chemical shift and NOE peak tables, to achieve nearly complete NOE assignments by CYANA. In other words, we have demonstrated that the idea would be feasible for samples that are suitable for a high-throughput NMR analysis with the highly automated strategy. The user is nevertheless advised to use some other validation tools such as ProCheck-NMR (Laskowski et al. 1996), WHATIF (Hooft et al. 1996; Rodriguez et al. 1998; Vriend 1990), QUEEN (Nabuurs et al. 2003), AVS (Moseley et al. 2004) or Protein-RPF (Huang et al. 2005), or tools working with additional experiments, such as residual dipolar coupling data, iDC (Wei and Werner 2006), all of which may help to avoid unforeseen problems in the assigned chemical shifts and the calculated structures.

Conclusion

Although CYANA is quite reliable to determine the correct structure, the few remaining faults in the chemical shifts and the NOE peak table must be addressed, in order to finalize the structure analysis. We developed a variety of modules, integrated in one software package which can seamlessly access the chemical shift table, the NOE assignment carried out by CYANA, and the calculated structure, and can greatly reduce the tediousness of the structure determination. These interactive modules implemented in KUJIRA are particularly useful to identify and correct the unassigned signals and the spurious NOEs that exist in the enormous amount of NMR data. We have established a new strategy using KUJIRA and CYANA, and have demonstrated its feasibility by an NMR structure study of a small protein by a non-expert, showing that accurate determinations of the chemical shifts and the structure can be achieved in a few weeks. The new strategy allows the NMR scientist to concentrate on monitoring the completeness of the NOE assignments and the structural problems in the NMR structure analysis, which will facilitate high-throughput studies by NMR.

Software availability

The KUJIRA software, installation instructions and examples are available at http://www.protein.gsc.riken.jp/Concept/kujira.html.

Acknowledgements We thank Drs. Toshio Yamazaki, Yutaka Muto, Fumiaki Hayashi, Takashi Nagata, Fahu He, Yufeng Wei and Prof. Milton, H. Werner for their valuable comments and discussions, and Prof. Yo Matsuo for his helpful advice regarding the algorithm for the calculation of the solvent accessible surface area. We also thank Mr. Masatomo Tanaka and Dr. Takuma Kasai for providing the information from the structural analyses of human ubiquitin. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. Curr Opin Struct Biol 14(5):547–553

Andrec M, Levy RM (2002) Protein sequential resonance assignments by combinatorial enumeration using 13C alpha chemical shifts and their (i, i-1) sequential connectivities. J Biomol NMR 23(4):263–270

Atreya HS, Sahu SC, Chary KV, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17(2):125–136

Baran MC, Moseley HN, Sahota G, Montelione GT (2002) SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. J Biomol NMR 24(2):113–121

Baran MC, Huang YJ, Moseley HN, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104(8):3541–3556

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6:1–10

Bartels C, Billeter M, Güntert P, Wüthrich K (1996a) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Bartels C, Billeter M, Güntert P, Wüthrich K (1996b) GARANT-A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18:139–149

Buchler NE, Zuiderweg ER, Wang H, Goldstein RA (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. J Magn Reson 125(1):34–42

Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 26(2):93–111

Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120(27):6836–6837

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13(3):289–302

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6(3):277–293

Fogh R, Ionides J, Ulrich E, Boucher W, Vranken W, Linge JP, Habeck M, Rieping W, Bhat TN, Westbrook J, Henrick K, Gilliland G, Berman H, Thornton J, Nilges M, Markley J, Laue E (2002) The CCPN project: an interim report on a data model for the NMR community. Nat Struct Biol 9(6):416–418

Fogh RH, Boucher W, Vranken WF, Pajon A, Stevens TJ, Bhat TN, Westbrook J, Ionides JM, Laue ED (2005) A framework for scientific data modeling and automated software development. Bioinformatics 21(8):1678–1684

Goddard TD, Kneller DG (2001) Sparkey 3. University of California, San Francisco

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 44:33–96

Gronwald W, Kirchfofer R, Gorler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR (1998) CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR 12:395–405

Güntert P (2003) Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 43:105–125

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273(1):283–298

Güntert P, Salzmann M, Braun D, Wüthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18(2):129–137

Helgstrand M, Kraulis P, Allard P, Hard T (2000) Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. J Biomol NMR 18(4):329–336

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319(1):209–227

Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381(6580):272

Huang YJ, Powers R, Montelione GT (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 127(6):1665–1674

Hyberts SG, Wagner G (2003) IBIS—a tool for automated sequential assignment of protein spectra from triple resonance experiments. J Biomol NMR 26(4):335–344

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genomics 4(2–3):179–189

Johnson BA, Blevins RA (1994) NMRView: a computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442(1):15–19

Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A, Yokoyama S (2004) Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. J Struct Funct Genomics 5(1–2):63–68

Kirby NI, DeRose EF, London RE, Mueller GA (2004) NvAssign: protein NMR spectral assignment with NMRView. Bioinformatics 20(7):1201–1203

Kraulis PJ (1989) ANSIG-a program for the assignment of protein H-1 2D NMR spectra by interactive computer graphics. J Magn Reson 24:627–633

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8(4):477–486

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379–400

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11(1):31–43

Li KB, Sanctuary BC (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. J Chem Inf Comput Sci 37(3):467–477

Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40(3):389–408

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (13C,15N)-labeled proteins. J Biomol NMR 9(2):151–166

Moseley HN, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR 28(4):341–355

Nabuurs SB, Spronk CA, Krieger E, Maassen H, Vriend G, Vuister GW (2003) Quantitative evaluation of experimental NMR restraints. J Am Chem Soc 125(39):12026–12034

Neidig KP, Geyer M, Görler A, Antz C, Saffrich R, Beneicke W, Kalbitzer HR (1995) Automated peak integration in multidimensional NMR spectra by an optimized iterative segmentation procedure. J Biol NMR 6:255–270

Nilges M, O'Donoghue SI (1998) Ambiguous NOEs and automated NOE assignment. Prog Nucl Magn Reson Spectrosc 32:107–139

Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G (1998) Homology modeling, model and software evaluation: three related resources. Bioinformatics 14(6):523–528

Slupsky CM, Boyko RF, Booth VK, Sykes BD (2003) Smartnotebook: a semi-automated approach to protein sequential NMR resonance assignments. J Biomol NMR 27(4):313–321

Spronk CA, Nabuurs SB, Krieger E, Vriend G, Vuister GW (2004) Validation of protein structures derived by NMR spectroscopy. Prog Nucl Magn Reson Spectrrosc 45(3–4):315–337

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59(4):687–696

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8(1):52–56, 29

Wei Y, Werner MH (2006) iDC: a comprehensive toolkit for the analysis of residual dipolar couplings for macromolecular structure determination. J Biomol NMR 35(1):17–25

Zimmerman D, Kulikowski C, Wang L, Lyons B, Montelione GT (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. J Biomol NMR 4(2):241–256

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269(4):592–610